Shivam Mokha

Data Science Case Study – Version 1

Analyze the attached dataset 'movie_metadata' and identify the insights which can be generated from this data. The dataset has 28 variables which are explained in detail in the 'variable description' tab.

Key Questions to be answered:

1. What are your observations based on exploration of this data?
2. What is the recipe to make a blockbuster, profitable movie? Share your hypothesis and insights based on the data here.

Callout – Considering the data as the whole dataset provided.

Answer 1:

Data cleaning:

The data provided consists of about 5043 rows of data and 28 columns. The data was further cleaned and completed as it consisted of multiple null values in various columns. The data was removed for about 48 duplicates.

Color column null -> Filled with the value 'Color' which is the more dominant category.
Director name column -> Filled with the value 'Unknown or multiple'.
Duration -> Filled with the mean value.
Actor 1, 2, 3 likes, no. of critic reviews, no. of user reviews, face on poster columns -> Filled with the value 0.
Language column-> Filled with the value 'English' which is the more dominant category.
Country column -> Filled with the value 'USA' which is the more dominant category.
Content rating column -> Filled with the value 'Unrated'.
Title year -> Since most values missing title year consisted of series and documentaries having no budget or gross, they were removed.
Aspect ratio column was dropped as it was not necessary and did not contribute to any insights.
Gross and Budget columns-> A web scrapper is used to extract data from IMDB to get missing budget and gross values (in USD). Approximately 600 missing values were filled.
A new column **net** is created which is the difference between gross and budget.

Shivam Mokha

Data exploration:

To start with data exploration, we find that the data correlation matrix as follow:

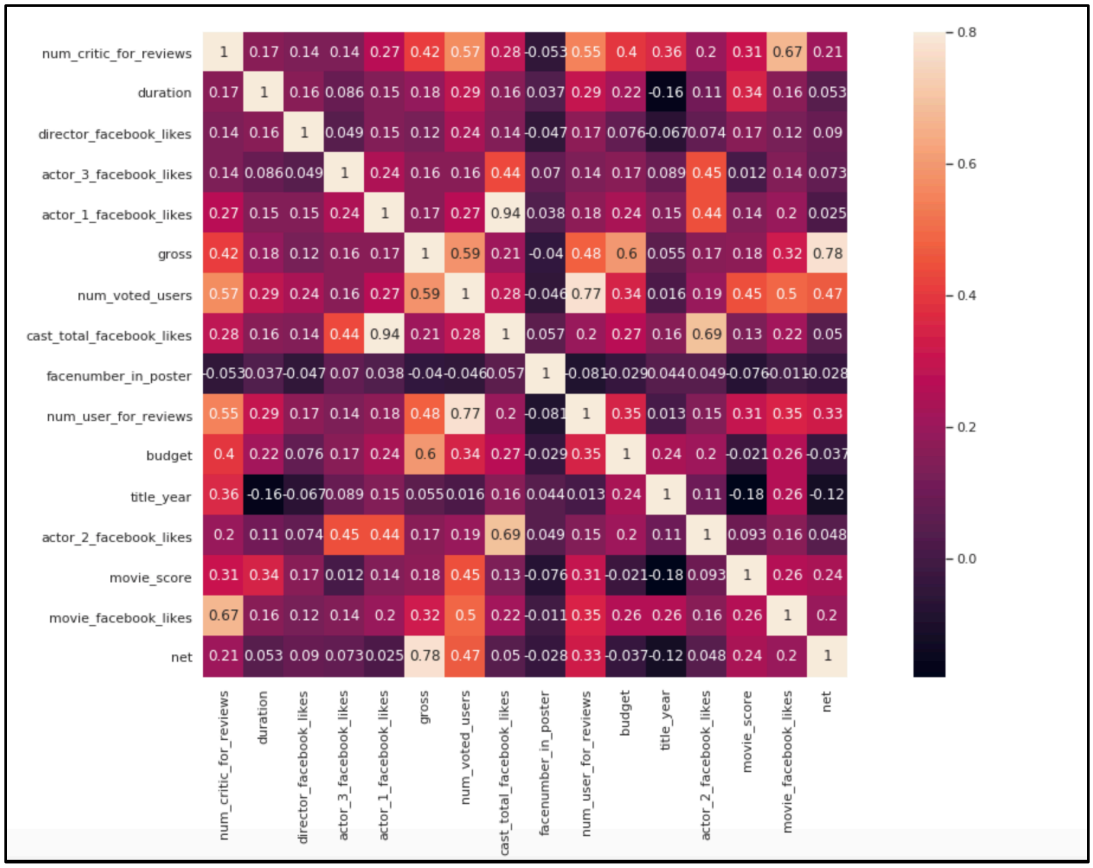

Fig. 1 Correlation heat map

- We find there is a high correlation between cast total Facebook likes and lead actor Facebook (0.94) likes which indicates very high influence of the primary actor over the whole caste.
- The correlation is further seen, with actor 2 Facebook like (0.69) and actor 3 Facebook like (0.44).
- We can also find the actor 2 and actor 3 have higher correlation with each other rather than with actor 1, which suggests a very high influence of actor 1 alone.
- Gross has a correlation with budget suggesting a big production movie is likely to sell higher tickets.

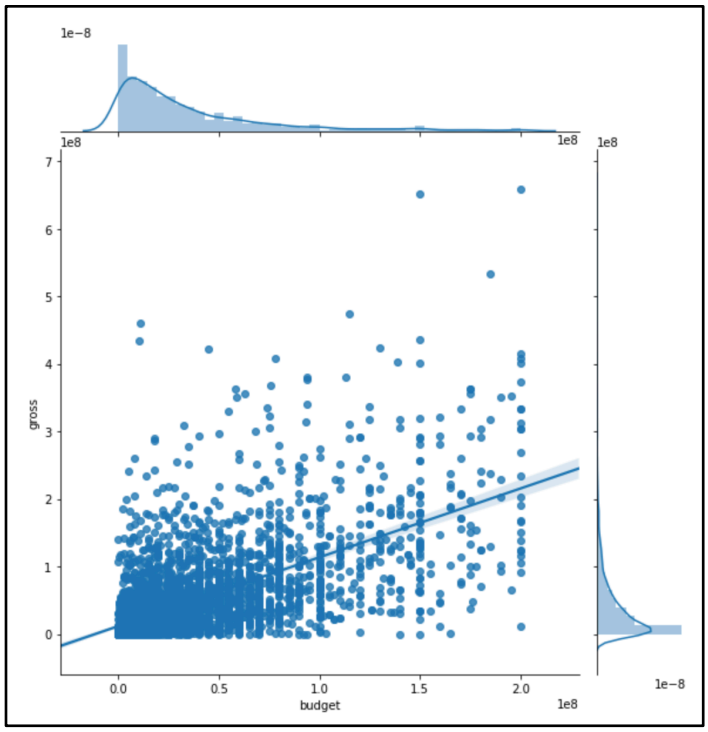

Fig. 2: Budget V/S Gross

- Gross is also highly corelated with ‘net’ (profit or loss) suggesting most of the with increase in ticket sales, profit is bound to increase.

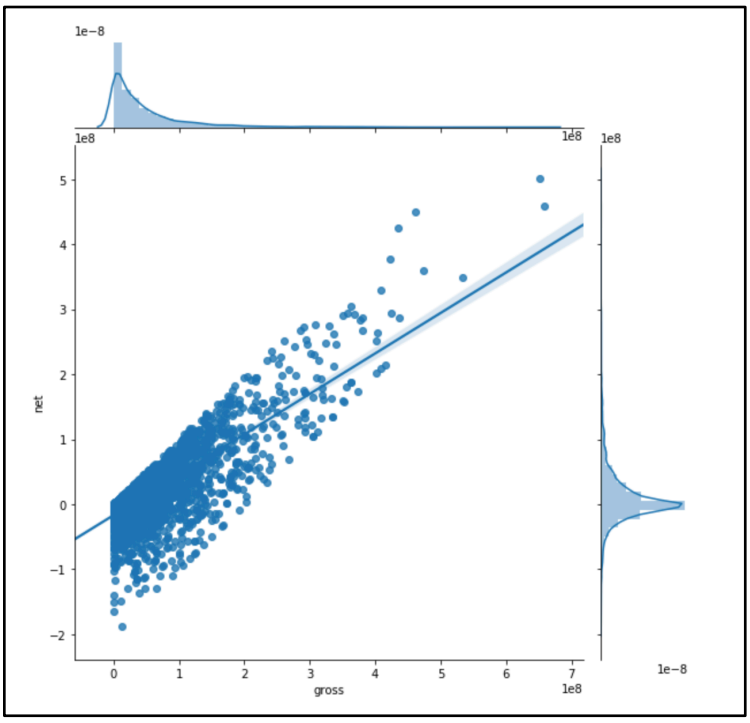

Fig. 3 Gross V/S Net

- Gross is also corelated to number of reviews by user (0.48) and critics (0.42) but more highly related with number of users (0.59) who voted for the movie.
- We find a weak correlation between gross and movie score which suggests that a movie might not have a very critically regarded to be successful.
- We find almost no correlation of number of actors on face cover with any of the other features.
- Plotting genre and gross, we find that a very high amount of average gross earnings is from 'Comedy' genre followed by 'Drama' and Family' movies. The least average earnings are from news, film-noir and short films. The same trend is followed by budget and net earnings.
- Considering a relation plot of year, the movie was made and budget, we find that there has been a steady rise in financial budget for movies.

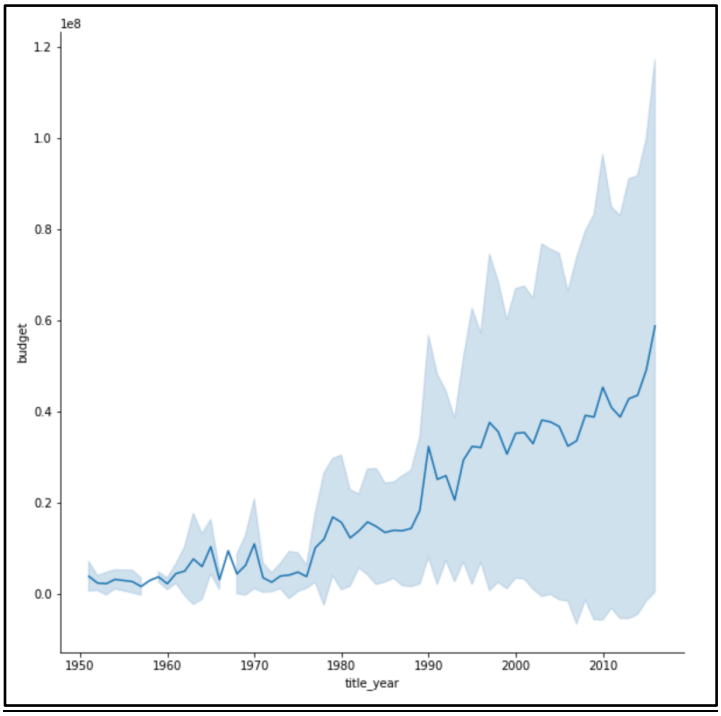

Fig. 6 Title Year V/S Budget

- Lastly, we find that the frequency of production of movies has increased over the past several years.

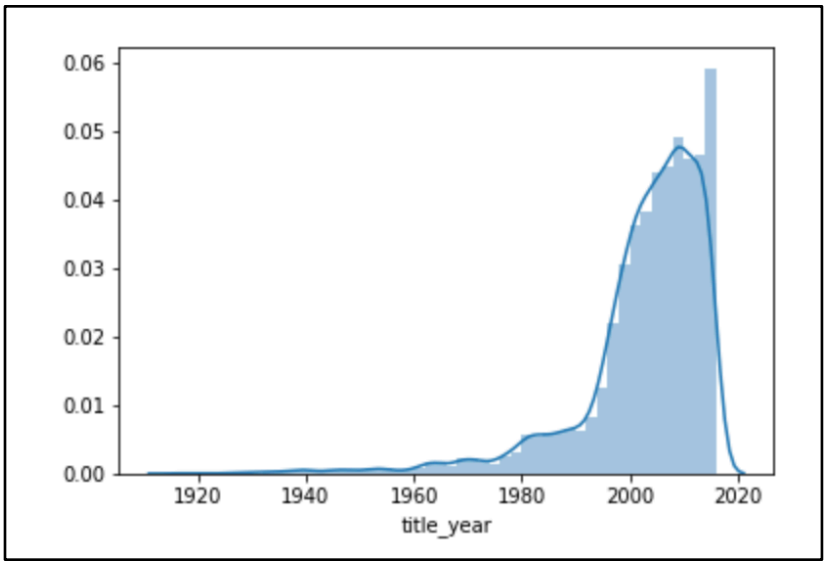

Fig 7. Title year distribution

Answer 2:

To make a blockbuster movie, we can consider the following for the hypothesis,

1. The director who made the movie has an impact on sales of the movie.
2. The director who has a higher count of movie productions and higher likes on Facebook has better financial returns.
3. Actor 1 has a higher impact on net returns as compared to actor 2 and actor 3.
4. Movie score does not have an impact on net earnings from the movie.
5. Quantity of people who vote, review or critique a movie has a direct impact on gross for the movie.
6. There has been a change in genre with respect to financial budget, gross and net earnings of a movie.

Insights:

Note:
For a movie to be considered a blockbuster profitable movie, we assume the following,

1. For a movie with a budget less than $5million, gross over $2million and having a profit greater than 500% of net
2. For a movie with a budget more than $5mil having a profit greater than 150% of net

If these conditions are satisfied, the blockbuster column is marked as 1, else 0

We find that Steven Spielberg has highest net earnings (over $2 billion) as well as highest gross of over $4 billion and has directed over 26 movies. Going down on the list, we find that even though George Lucas and Joss Whedon have done lower amount of movies, they have a higher average net earnings than Steven Spielberg.

Shivam Mokha

Director Name	Count of Movies Done	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net
Steven Spielberg	26	$1,662,900,870.00	$4,145,988,843.00	$2,483,087,973.00	149.32%	$95,503,383.58
George Lucas	5	$354,777,000.00	$1,741,418,480.00	$1,386,641,480.00	390.85%	$277,328,296.00
James Cameron	7	$748,500,000.00	$1,948,125,910.00	$1,199,625,910.00	160.27%	$171,375,130.00
Joss Whedon	4	$730,000,000.00	$1,730,886,628.00	$1,000,886,628.00	137.11%	$250,221,657.00
Chris Columbus	11	$677,000,000.00	$1,618,707,624.00	$941,707,624.00	139.10%	$85,609,784.00
Peter Jackson	12	$1,692,000,000.00	$2,592,969,279.00	$900,969,279.00	53.25%	$75,080,773.25
Tim Burton	16	$1,247,000,000.00	$2,071,275,480.00	$824,275,480.00	66.10%	$51,517,217.50
Christopher Nolan	8	$1,005,000,000.00	$1,813,227,576.00	$808,227,576.00	80.42%	$101,028,447.00
Jon Favreau	7	$793,000,000.00	$1,562,381,547.00	$769,381,547.00	97.02%	$109,911,649.57
Francis Lawrence	5	$603,000,000.00	$1,358,501,971.00	$755,501,971.00	125.29%	$151,100,394.20

Table 1: Sorted by net earnings

Director Name	Count of Movies Done	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net
Steven Spielberg	26	$1,662,900,870.00	$4,145,988,843.00	$2,483,087,973.00	149.32%	$95,503,383.58
Peter Jackson	12	$1,692,000,000.00	$2,592,969,279.00	$900,969,279.00	53.25%	$75,080,773.25
Michael Bay	13	$1,587,000,000.00	$2,231,242,537.00	$644,242,537.00	40.59%	$49,557,118.23
Tim Burton	16	$1,247,000,000.00	$2,071,275,480.00	$824,275,480.00	66.10%	$51,517,217.50
Sam Raimi	12	$1,430,600,000.00	$2,049,549,198.00	$618,949,198.00	43.27%	$51,579,099.83
James Cameron	7	$748,500,000.00	$1,948,125,910.00	$1,199,625,910.00	160.27%	$171,375,130.00
Christopher Nolan	8	$1,005,000,000.00	$1,813,227,576.00	$808,227,576.00	80.42%	$101,028,447.00
George Lucas	5	$354,777,000.00	$1,741,418,480.00	$1,386,641,480.00	390.85%	$277,328,296.00
Joss Whedon	4	$730,000,000.00	$1,730,886,628.00	$1,000,886,628.00	137.11%	$250,221,657.00
Robert Zemeckis	13	$1,070,000,000.00	$1,619,309,108.00	$549,309,108.00	51.34%	$42,254,546.77

Table 2: Sorted by gross earnings

Going by count of movies, we find that it does not affect gross and net earnings, which is evident from directors Ridley Scott and Renny Harlin who have incurred loss over very high number of movies.

Director Name	Count of Movies Done	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net
Steven Spielberg	26	$1,662,900,870.00	$4,145,988,843.00	$2,483,087,973.00	149.32%	$95,503,383.58
Woody Allen	21	$312,500,000.00	$344,706,462.00	$32,206,462.00	10.31%	$1,533,641.05
Clint Eastwood	20	$773,600,000.00	$1,394,021,100.00	$620,421,100.00	80.20%	$31,021,055.00
Martin Scorsese	18	$929,600,000.00	$959,272,374.00	$29,672,374.00	3.19%	$1,648,465.22
Ridley Scott	17	$1,430,000,000.00	$1,337,771,688.00	-$92,228,312.00	-6.45%	-$5,425,194.82
Tim Burton	16	$1,247,000,000.00	$2,071,275,480.00	$824,275,480.00	66.10%	$51,517,217.50
Steven Soderbergh	16	$643,200,000.00	$1,050,729,636.00	$407,529,636.00	63.36%	$25,470,602.25
Spike Lee	15	$249,575,000.00	$328,500,427.00	$78,925,427.00	31.62%	$5,261,695.13
Renny Harlin	15	$692,300,000.00	$523,975,947.00	-$168,324,053.00	-24.31%	-$11,221,603.53
Oliver Stone	14	$584,600,000.00	$681,292,713.00	$96,692,713.00	16.54%	$6,906,622.36

Table 3: Sort by movie count

Taking into account director likes on Facebook, we find that directors have higher count of movie production as well as number of users who have voted for them. We also find that they have an average movie score around 7.

Shivam Mokha

Director Name	Count of Movies Done	Net Profit Percentage	Average of Net	Director Facebook Likes	No. of users voted	Average movie score
Steven Spielberg	26	149.32%	$95,503,383.58	364000	8298294	7.48
Clint Eastwood	20	80.20%	$31,021,055.00	320000	3066098	7.23
Martin Scorsese	18	3.19%	$1,648,465.22	306000	5485696	7.66
Woody Allen	21	10.31%	$1,533,641.05	231000	1566131	7.00
David Fincher	10	32.91%	$21,390,452.40	210000	5185902	7.75
Tim Burton	16	66.10%	$51,517,217.50	208000	3502256	6.93
Christopher Nolan	8	80.42%	$101,028,447.00	176000	8106282	8.43
Tony Scott	12	21.48%	$12,996,139.42	144000	1581167	6.79
Quentin Tarantino	8	107.91%	$40,492,449.25	128000	5590660	8.20
Harold Ramis	8	47.97%	16705296.75	88000	966874	6.55

Table 4: Sort by Facebook likes

Taking on movie scores into account we find no correlation between movies scores and high net earnings.

Director Name	Count of Movies Done	Net Profit Percentage	Average of Net	Director Facebook Likes	No. of users voted	Average movie score
Cary Bell	1	-92.06%	-$165,703.00	0	27	8.70
Tony Kaye	1	-10.50%	-$787,759.00	194	782437	8.60
Michael Curtiz	1	7.85%	$74,560.00	345	387508	8.60
Charles Chaplin	1	-89.12%	-$1,336,755.00	0	143086	8.60
Ron Fricke	1	-34.95%	-$1,398,153.00	330	22457	8.50
Raja Menon	1	536.36%	$23,600,000.00	6	30977	8.50
Majid Majidi	1	414.11%	$745,402.00	373	27882	8.50
Damien Chazelle	1	296.73%	$9,792,000.00	141	399138	8.50
Sergio Leone	4	-44.45%	-$4,044,623.00	0	1088080	8.48
Christopher Nolan	8	80.42%	101028447	176000	8106282	8.43
S.S. Rajamouli	1	-63.95%	-11528148	50	62756	8.40

Table 5: Sort by movie scores

Looking at net earnings by profit, we can see that highest net earnings for actor 1 is greater than that of actor 2 and actor 3 suggesting a higher impact on movie earnings from actor 1.It can also be seen that the earnings for actor 2 and actor 3 is more comparable with each other.

Actor_1_Name	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net	Facebook Likes
Harrison Ford	$1,306,377,000.00	$3,403,206,163.00	$2,096,829,163.00	160.51%	$77,660,339.37	297000.00
Tom Hanks	$1,696,000,000.00	$3,264,559,076.00	$1,568,559,076.00	92.49%	$65,356,628.17	360000.00
Jennifer Lawrence	$1,182,150,000.00	$2,367,856,024.00	$1,185,706,024.00	100.30%	$84,693,287.43	476000.00
Tom Cruise	$1,838,800,000.00	$2,988,868,140.00	$1,150,068,140.00	62.54%	$46,002,725.60	250000.00
Robert Pattinson	$712,000,000.00	$1,841,497,127.00	$1,129,497,127.00	158.64%	$141,187,140.88	168000.00
Bradley Cooper	$609,800,000.00	$1,658,315,287.00	$1,048,515,287.00	171.94%	$80,655,022.08	182000.00
Chris Hemsworth	$1,735,000,000.00	$2,729,132,988.00	$994,132,988.00	57.30%	$66,275,532.53	390000.00
Robin Williams	$1,347,600,000.00	$2,296,971,817.00	$949,371,817.00	70.45%	$36,514,300.65	1274000.00
J.K. Simmons	$1,907,300,000.00	$2,856,407,143.00	$949,107,143.00	49.76%	$30,616,359.45	744000.00
Leonardo DiCaprio	$1,731,500,000.00	$2,640,581,774.00	$909,081,774.00	52.50%	$43,289,608.29	609000.00

Actor_2_Name	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net	Facebook Likes
Robert Downey Jr.	$690,000,000.00	$1,705,550,693.00	$1,015,550,693.00	147.18%	$338,516,897.67	63000.00
John Ratzenberger	$770,000,000.00	$1,586,229,220.00	$816,229,220.00	106.00%	$136,038,203.33	6000.00
Kristen Stewart	$528,000,000.00	$1,336,856,408.00	$808,856,408.00	153.19%	$161,771,281.60	85000.00
Orlando Bloom	$1,157,000,000.00	$1,872,766,986.00	$715,766,986.00	61.86%	$102,252,426.57	35000.00
Vin Diesel	$492,000,000.00	$1,193,764,891.00	$701,764,891.00	142.64%	$100,252,127.29	98000.00
Judy Greer	$565,500,000.00	$1,198,610,442.00	$633,110,442.00	111.96%	$70,345,604.67	18000.00
Stephen Root	$298,000,000.00	$909,501,279.00	$611,501,279.00	205.20%	$76,437,659.88	7512.00
Josh Hutcherson	$273,000,000.00	$860,690,372.00	$587,690,372.00	215.27%	$195,896,790.67	42000.00
James Franco	$1,125,200,000.00	$1,701,867,383.00	$576,667,383.00	51.25%	$52,424,307.55	121000.00
Joel David Moore	$312,000,000.00	$840,479,166.00	$528,479,166.00	169.38%	$176,159,722.00	2808.00

Actor_3_Name	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net	Facebook Likes
Scarlett Johansson	$730,000,000.00	$1,758,633,436.00	$1,028,633,436.00	140.91%	$257,158,359.00	76000.00
Kenny Baker	$48,500,000.00	$812,426,674.00	$763,926,674.00	1575.11%	$254,642,224.67	1512.00
Steve Coogan	$687,000,000.00	$1,393,134,638.00	$706,134,638.00	102.79%	$88,266,829.75	8000.00
Anna Kendrick	$192,000,000.00	$811,419,412.00	$619,419,412.00	322.61%	$154,854,853.00	40000.00
Kirsten Dunst	$953,000,000.00	$1,562,368,932.00	$609,368,932.00	63.94%	$87,052,704.57	28000.00
Taylor Lautner	$290,000,000.00	$881,221,480.00	$591,221,480.00	203.87%	$197,073,826.67	36000.00
Omar Sy	$150,000,000.00	$652,177,271.00	$502,177,271.00	334.78%	$502,177,271.00	1000.00
Wes Studi	$312,000,000.00	$784,364,349.00	$472,364,349.00	151.40%	$157,454,783.00	2565.00
Billy Boyd	$187,000,000.00	$654,316,475.00	$467,316,475.00	249.90%	$233,658,237.50	1714.00
Hayden Christensen	$228,000,000.00	$690,938,138.00	$462,938,138.00	203.04%	$231,469,069.00	8000.00

Table 6: Sort by net earnings

It can also be seen that actor 1 Facebook likes are higher than actor 2 and actor 3 Facebook likes. Also, the top 10 highest actors 1,2 and 3 with highest Facebook likes seem to have high budgets suggesting higher wage demands by these actors.

Actor_1_Name	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net	Facebook Likes
Johnny Depp	$3,136,600,000.00	$3,714,788,902.00	$578,188,902.00	18.43%	$14,825,356.46	1560000.00
Robin Williams	$1,347,600,000.00	$2,296,971,817.00	$949,371,817.00	70.45%	$36,514,300.65	1274000.00
Robert De Niro	$1,599,800,000.00	$2,199,151,420.00	$599,351,420.00	37.46%	$12,752,157.87	1034000.00
J.K. Simmons	$1,907,300,000.00	$2,856,407,143.00	$949,107,143.00	49.76%	$30,616,359.45	744000.00
Jason Statham	$1,001,474,000.00	$1,111,401,945.00	$109,927,945.00	10.98%	$4,397,117.80	650000.00
Darcy Donavan	$26,000,000.00	$84,136,909.00	$58,136,909.00	223.60%	$58,136,909.00	640000.00
Leonardo DiCaprio	$1,731,500,000.00	$2,640,581,774.00	$909,081,774.00	52.50%	$43,289,608.29	609000.00
Jimmy Bennett	$498,000,000.00	$329,550,504.00	-$168,449,496.00	-33.83%	-$24,064,213.71	609000.00
Robert Downey Jr.	$1,717,000,000.00	$2,456,990,061.00	$739,990,061.00	43.10%	$28,461,156.19	546000.00
Denzel Washington	$1,454,500,000.00	$1,896,762,381.00	$442,262,381.00	30.41%	$14,742,079.37	540000.00

Actor_2_Name	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net	Facebook Likes
Morgan Freeman	$1,082,400,000.00	$1,419,284,244.00	$336,884,244.00	31.12%	$16,844,212.20	220000.00
Brad Pitt	$921,000,000.00	$1,046,946,193.00	$125,946,193.00	13.67%	$8,996,156.64	154000.00
Andrew Fiscella	$40,000,000.00	$66,466,372.00	$26,466,372.00	66.17%	$26,466,372.00	137000.00
Charlize Theron	$863,000,000.00	$840,198,843.00	-$22,801,157.00	-2.64%	-$1,628,654.07	126000.00
Meryl Streep	$386,000,000.00	$637,692,181.00	$251,692,181.00	65.21%	$22,881,107.36	121000.00
James Franco	$1,125,200,000.00	$1,701,867,383.00	$576,667,383.00	51.25%	$52,424,307.55	121000.00
Bruce Willis	$493,000,000.00	$454,336,712.00	-$38,663,288.00	-7.84%	-$4,295,920.89	117000.00
Christian Bale	$420,000,000.00	$768,850,335.00	$348,850,335.00	83.06%	$69,770,067.00	115000.00
Kate Winslet	$499,500,000.00	$882,170,851.00	$382,670,851.00	76.61%	$47,833,856.38	112000.00
Robert De Niro	$174,000,000.00	$248,192,939.00	$74,192,939.00	42.64%	$14,838,587.80	110000.00

Actor_3_Name	Sum of Budget	Sum of Gross	Sum of Net	Net Profit Percentage	Average of Net	Facebook Likes
Anne Hathaway	$777,000,000.00	$1,217,808,439.00	$440,808,439.00	56.73%	$62,972,634.14	77000.00
Scarlett Johansson	$730,000,000.00	$1,758,633,436.00	$1,028,633,436.00	140.91%	$257,158,359.00	76000.00
Gary Oldman	$433,000,000.00	$402,839,726.00	-$30,160,274.00	-6.97%	-$6,032,054.80	50000.00
Joseph Gordon-Levitt	$410,000,000.00	$740,699,493.00	$330,699,493.00	80.66%	$165,349,746.50	46000.00
Steve Carell	$386,000,000.00	$392,084,113.00	$6,084,113.00	1.58%	$1,014,018.83	42000.00
Bradley Cooper	$121,000,000.00	$338,650,199.00	$217,650,199.00	179.88%	$72,550,066.33	42000.00
Anna Kendrick	$192,000,000.00	$811,419,412.00	$619,419,412.00	322.61%	$154,854,853.00	40000.00
Taylor Lautner	$290,000,000.00	$881,221,480.00	$591,221,480.00	203.87%	$197,073,826.67	36000.00
Matthew McConaughey	$103,000,000.00	$200,945,952.00	$97,945,952.00	95.09%	$32,648,650.67	33000.00
James Franco	$479,900,000.00	$491,363,392.00	$11,463,392.00	2.39%	$3,821,130.67	33000.00

Table 7: Sort by Facebook Likes

Considering the highest earning movies, we find that these movies have a high number of users and critics who have voted and reviewed for them respectively, suggesting an influence of users and critics. These movies usually have an average rating of about 7 or more. It should also be noted that the movies with highest average scores are not high on net earnings which is possible due to in influx of audience of movies in the recent past and which was not present before

Movie	Budget	Gross	Net	Net %	User Reviews	Users Voted	Movie Score	Critic Reviews
The Avengers	$440,000,000.00	$1,246,559,094.00	$806,559,094.00	1.83308885	3444	1990830	8.1	1406
Avatar	$237,000,000.00	$760,505,847.00	$523,505,847.00	2.20888543	3054	886204	7.9	723
Jurassic World	$150,000,000.00	$652,177,271.00	$502,177,271.00	3.347848473	1290	418214	7	644
Titanic	$200,000,000.00	$658,672,302.00	$458,672,302.00	2.29336151	2528	793059	7.7	315
Star Wars: Episode IV - A New Hope	$11,000,000.00	$460,935,665.00	$449,935,665.00	40.90324227	1470	911097	8.7	282
E.T. the Extra-Terrestrial	$10,500,000.00	$434,949,459.00	$424,449,459.00	40.423758	515	281842	7.9	215
The Lion King	$45,000,000.00	$422,783,777.00	$377,783,777.00	8.395195044	656	644348	8.5	186
The Jungle Book	$350,000,000.00	$725,290,282.00	$375,290,282.00	1.072257949	796	212293	7.8	740
Star Wars: Episode I - The Phantom Menace	$115,000,000.00	$474,544,677.00	$359,544,677.00	3.126475452	3597	534658	6.5	320
The Dark Knight	$185,000,000.00	$533,316,061.00	$348,316,061.00	1.882789519	4667	1676169	9	645

Table 8: Sort by net earnings

With respect to genre, the highest earning genre has high number of users who have voted for it along with high number of critics and user reviews. The genres with highest chance of success is ‘Music’ followed by ‘Horror’. ‘Crime’, ‘History’ and ‘Action’ movies have a relatively low chance of success.

Shivam Mokha

Genre	Count of genres	Average of budget	Sum of gross	Net % val	Average of duration	Sum of num_user_for_reviews	Sum of num_voted_users	Sum of num_critic_for_reviews	Average of blockbuster
Comedy	1590	$32,467,452.08	$76,653,264,404.00	48.49%	100.9113208	327749	114410824	201730	20.94%
Drama	2098	$27,063,244.72	$74,181,935,971.00	30.65%	115.8522402	620208	192147945	320102	18.73%
Adventure	815	$68,498,087.02	$71,567,314,625.00	28.20%	111.5852761	345491	115522250	151810	16.69%
Action	987	$58,888,285.47	$67,630,452,729.00	16.36%	111.7264438	410802	130600787	181927	13.27%
Thriller	1187	$38,206,822.95	$54,358,966,081.00	19.86%	110.3142376	433943	127481219	216166	15.75%
Fantasy	536	$57,771,566.65	$41,262,704,360.00	33.25%	104.7369403	200404	64588825	93753	18.66%
Family	475	$60,802,638.80	$40,811,215,600.00	41.31%	98.48421053	102024	40064744	64092	20.42%
Romance	938	$28,411,147.71	$38,552,192,780.00	44.66%	110.322176	229884	68486555	121840	20.36%
Sci-Fi	518	$58,014,527.79	$37,740,351,327.00	25.59%	109.003861	259748	79453706	109773	15.25%
Crime	750	$31,238,491.41	$28,364,206,834.00	21.07%	111.0853333	216214	79482226	115469	14.53%
Animation	203	$78,037,832.51	$21,677,745,775.00	36.84%	90.54679803	45728	23245465	34181	16.26%
Mystery	416	$33,592,707.51	$18,049,177,057.00	29.16%	110.6514423	161916	46604667	79383	19.23%
Horror	438	$20,102,703.22	$14,190,489,468.00	61.16%	100.5	163103	32229778	82977	29.91%
Biography	260	$25,790,022.25	$9,171,971,445.00	36.78%	125.6153846	66254	24449297	45305	20.38%
War	183	$36,569,288.69	$6,963,537,903.00	4.05%	132.726776	62031	17856556	25817	15.30%
Sport	162	$31,438,026.98	$6,886,319,733.00	35.21%	111.5740741	30149	10106892	19671	22.22%
Music	172	$21,637,804.35	$6,352,249,685.00	70.68%	109.5872093	30353	7841438	19311	30.81%
History	181	$36,394,694.38	$6,137,133,824.00	-6.84%	136.9668508	52640	14954110	27646	13.81%
Musical	110	$30,309,781.95	$5,556,662,623.00	66.66%	109.1272727	28134	6748606	12529	20.00%
Western	78	$31,527,789.76	$2,709,455,121.00	10.18%	124.5384615	19058	6285158	9742	16.67%
Documentary	72	$6,167,035.13	$932,718,179.00	110.06%	95.58333333	8129	1240348	5775	23.61%
Film-Noir	4	$1,633,594.25	$20,827,927.00	218.74%	104.5	748	164546	428	25.00%
Short	2	$3,017,000.00	$7,852,534.00	30.14%	38	7	269	10	0.00%
News	1	$1,500,000.00	$124,244.00	-91.72%	108	42	6678	68	0.00%
Grand Total	**12176**	**$39,532,788.62**	**$629,778,870,229.00**	**30.84%**	**109.8999837**	**3814759**	**1203972889**	**1939505**	**18.46%**

Thus, to conclude, we can find that for a successful blockbuster movie, we should look at the following,

1. Direct with high counts of movies along with high average movie ratings (about 7) and Facebook likes.
2. Lead actor (actor 1) with high amount of Facebook followed along with actor 2 and actor 3. Actor 1 can be compensated with better selection of actor 2 and actor 3.
3. Attract high number of users and reviews could impact the gross for movie sales.
4. To increase chances of success, movies from genres like ‘Music, ‘Comedy’ or ‘Sports’ could be produced.

The python notebook also contains a logistic regression which splits the test and train set in the ratio 3:7. The accuracy of the best model is 0.97 and can explain about 38.5% of the data with the model. Further, PCA can be used to reduce number of features to 3 but would reduce the accuracy to 0.83

```
In [673]: predictions = logreg.predict(x_test)

In [674]: score = logreg.score(x_test, y_test)
          print(score)

          0.9753636790239324

In [675]: cm = metrics.confusion_matrix(y_test, predictions)
          print(cm)

          [[6910   60]
           [ 150 1404]]

In [678]: X2 = sm.add_constant(x)
          est = sm.OLS(y, X2)
          est2 = est.fit()
          print(est2.summary())
```

OLS Regression Results

Dep. Variable:	blockbuster	R-squared:	0.389
Model:	OLS	Adj. R-squared:	0.385
Method:	Least Squares	F-statistic:	98.68
Date:	Thu, 04 Jul 2019	Prob (F-statistic):	0.00
Time:	18:45:23	Log-Likelihood:	-2751.4
No. Observations:	12176	AIC:	5661.
Df Residuals:	12097	BIC:	6246.
Df Model:	78		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1608	0.021	7.576	0.000	0.119	0.202
num_critic_for_reviews	0.0003	4.86e-05	6.154	0.000	0.000	0.000
director_facebook_likes	-2.26e-06	9.97e-07	-2.268	0.023	-4.21e-06	-3.07e-07
actor_3_facebook_likes	1.892e-06	4.63e-06	0.409	0.683	-7.18e-06	1.1e-05
actor_1_facebook_likes	4.478e-06	3.03e-06	1.478	0.139	-1.46e-06	1.04e-05

Fig. 8 Best model