

EXPLORATORY DATA ANAYSIS REPORT ON THE MOVIE INDUSTRY



TABLE OF CONTENTS

03

BACKGROUND

04

DATA CLEANING &
PREPARATION

11

DESCRIPTIVE
ANALYSIS

13

CORRELATION
ANALYSIS

16

VISUALIZATION

23

RECOMMENDATIONS



BACKGROUND

Cinema is an art form that weaves a narrative into a visual and auditory experience, thus imprinting an indelible mark on our collective consciousness. There is no single "most important" element as all film components — story, characters, acting, direction, production quality, pacing, emotional impact, relatability, and originality — contribute to its overall quality. But what makes a good movie? This question has stirred up numerous debates among film critics and movie buffs. Let's embark on a data analytics journey through the various key factors that the dataset at our disposal is made up of. To analyze and visualize the given data, we can focus on extracting insights such as the distribution of movie budgets, IMDB scores, genres, gross earnings, and other factors like the relationship between budget and gross, or the impact of an actor's or director's social media presence on movie success.

Here's a step-by-step breakdown of how we can approach this: data cleaning and preparation, descriptive statistics, correlation analysis, visualization, specific information extraction and data insights.

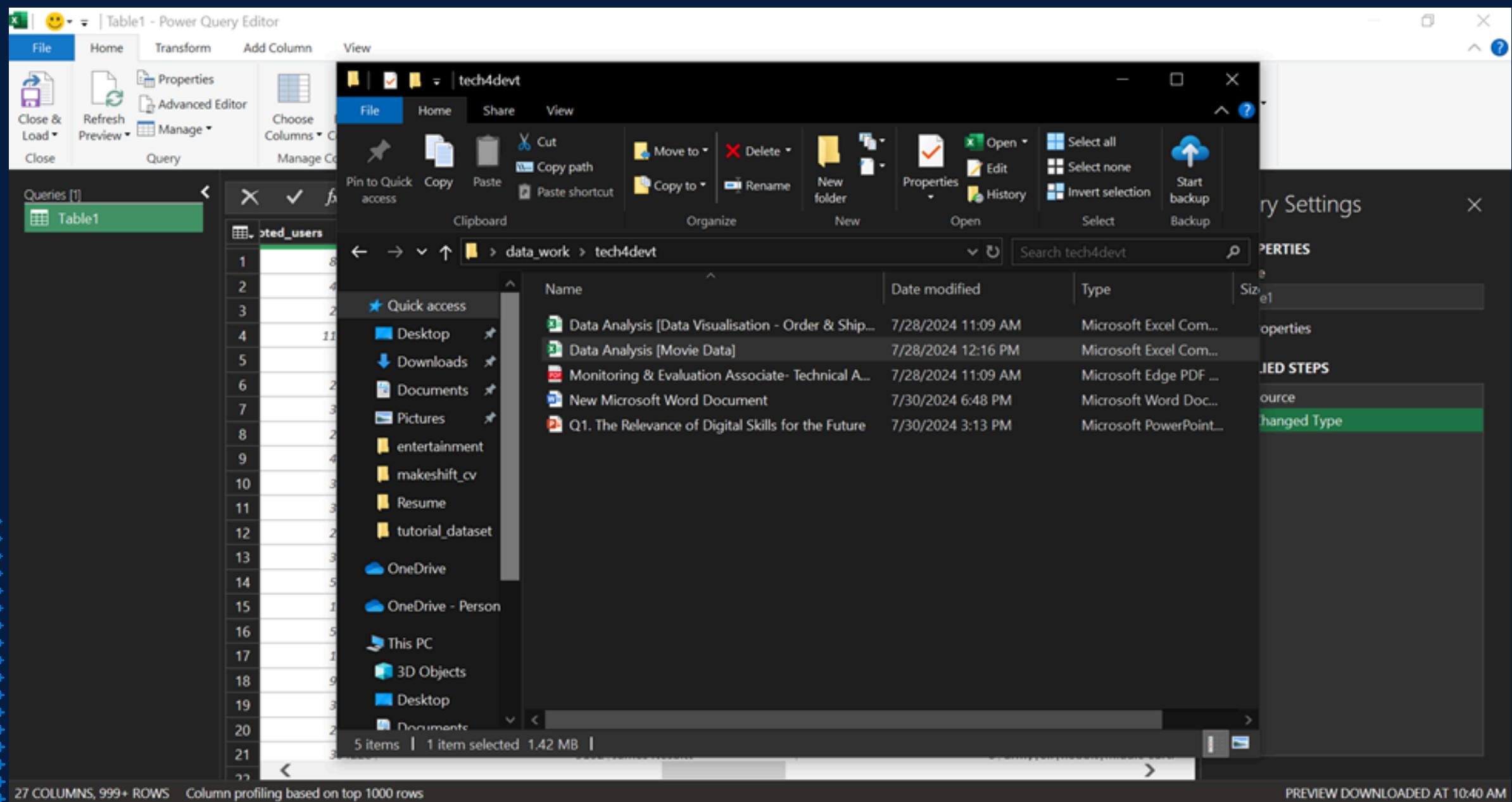


DATA PREPARATION & CLEANING

Data collection

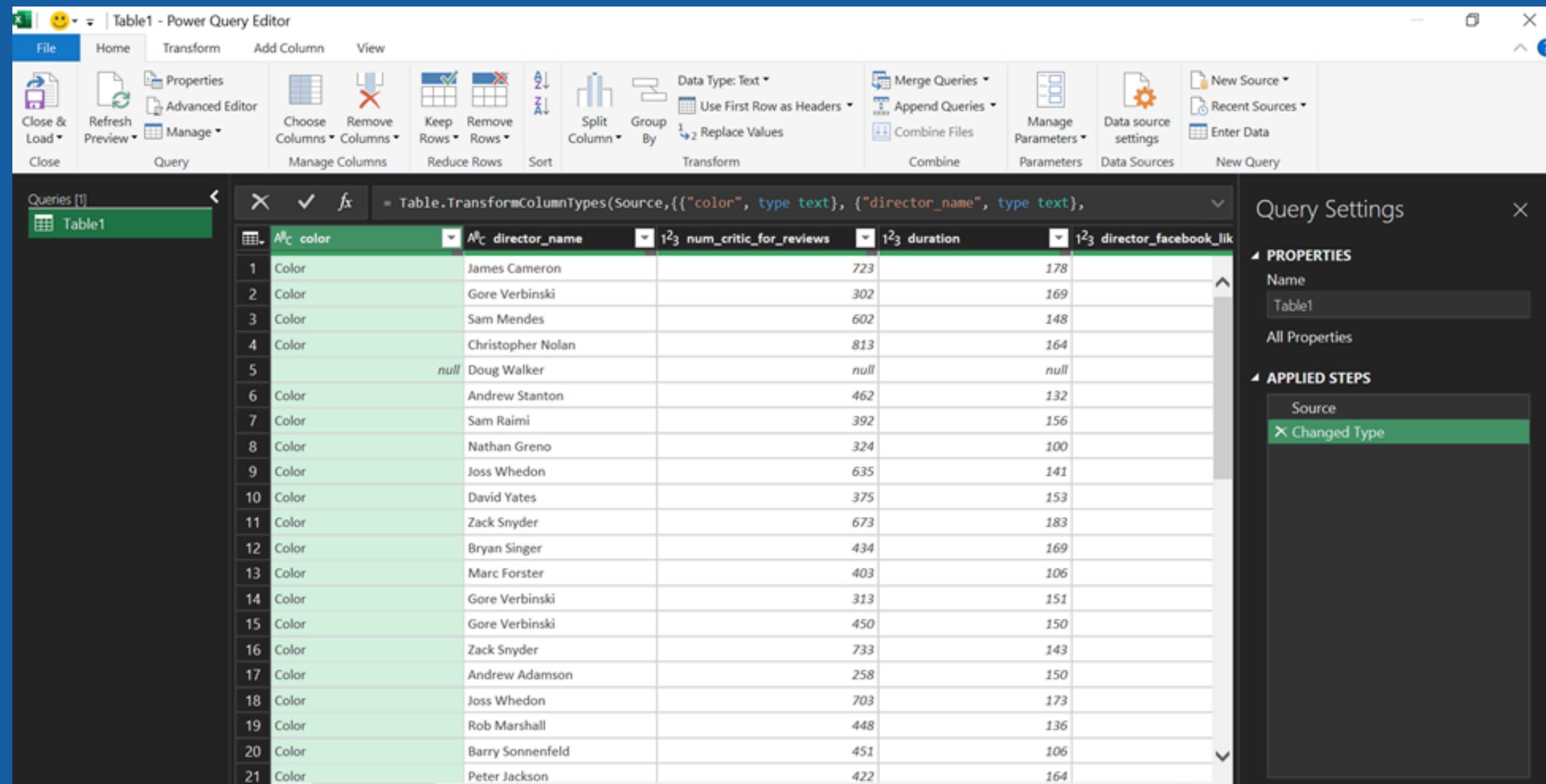
The dataset for this project was sourced from the following:

c:\Users\USER\Desktop\data_work\tech4devt



Dataset migration

The dataset was then imported from CSV file format into Excel Power Query

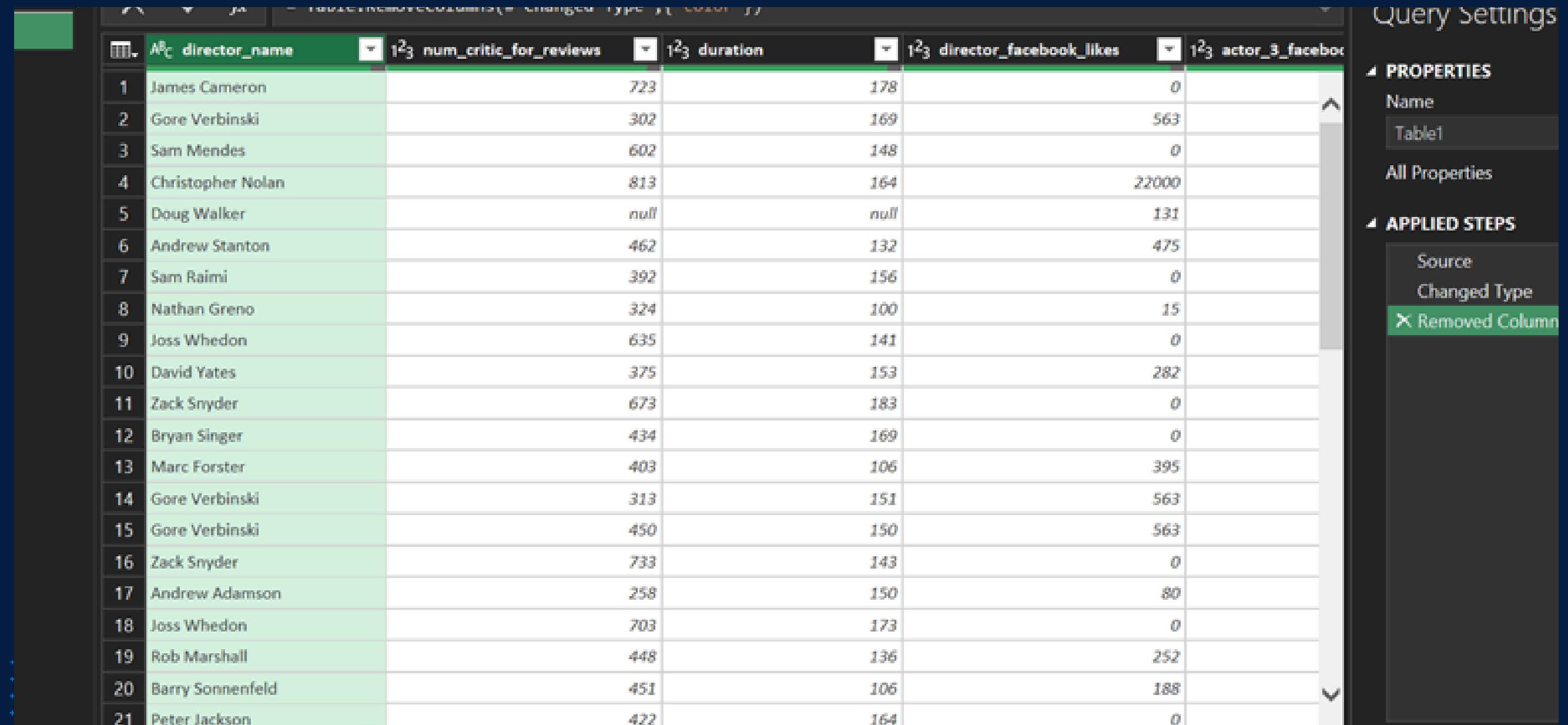


The screenshot shows the Microsoft Power Query Editor interface. The main area displays a table named "Table1" with 21 rows and 5 columns. The columns are: "color", "director_name", "num_critic_for_reviews", "duration", and "director_facebook_likes". The "color" column contains values like "Color" and "null". The "director_name" column contains names like "James Cameron", "Gore Verbinski", "Sam Mendes", etc. The "num_critic_for_reviews" column contains numerical values like 723, 302, 602, etc. The "duration" column contains numerical values like 178, 169, 148, etc. The "director_facebook_likes" column contains numerical values like 106, 136, 150, etc. The "APPLIED STEPS" pane on the right shows a single step: "Changed Type".

color	director_name	num_critic_for_reviews	duration	director_facebook_likes
Color	James Cameron	723	178	106
Color	Gore Verbinski	302	169	136
Color	Sam Mendes	602	148	150
Color	Christopher Nolan	813	164	173
null	Doug Walker	null	null	106
Color	Andrew Stanton	462	132	136
Color	Sam Raimi	392	156	150
Color	Nathan Greno	324	100	106
Color	Joss Whedon	635	141	173
Color	David Yates	375	153	136
Color	Zack Snyder	673	183	173
Color	Bryan Singer	434	169	136
Color	Marc Forster	403	106	106
Color	Gore Verbinski	313	151	150
Color	Gore Verbinski	450	150	150
Color	Zack Snyder	733	143	173
Color	Andrew Adamson	258	150	150
Color	Joss Whedon	703	173	173
Color	Rob Marshall	448	136	136
Color	Barry Sonnenfeld	451	106	106
Color	Peter Jackson	422	164	173

Data Transformation

Removal of unwanted columns



	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes
1	James Cameron	723	178	0	
2	Gore Verbinski	302	169	563	
3	Sam Mendes	602	148	0	
4	Christopher Nolan	813	164	22000	
5	Doug Walker	null	null	131	
6	Andrew Stanton	462	132	475	
7	Sam Raimi	392	156	0	
8	Nathan Greno	324	100	15	
9	Joss Whedon	635	141	0	
10	David Yates	375	153	282	
11	Zack Snyder	673	183	0	
12	Bryan Singer	434	169	0	
13	Marc Forster	403	106	395	
14	Gore Verbinski	313	151	563	
15	Gore Verbinski	450	150	563	
16	Zack Snyder	733	143	0	
17	Andrew Adamson	258	150	80	
18	Joss Whedon	703	173	0	
19	Rob Marshall	448	136	252	
20	Barry Sonnenfeld	451	106	188	
21	Peter Jackson	422	164	0	

Query Settings

PROPERTIES

Name

Table1

All Properties

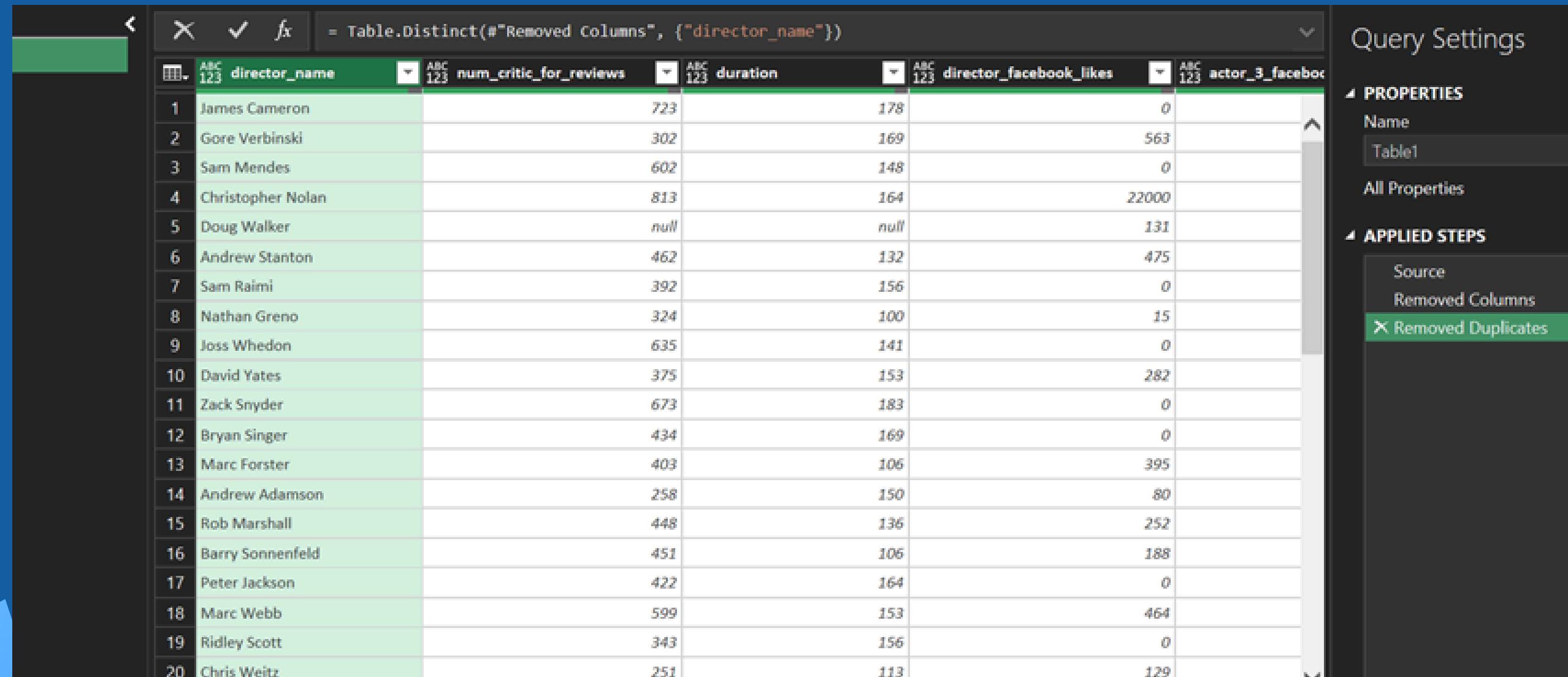
APPLIED STEPS

Source

Changed Type

Removed Column

Removal of duplicates



The screenshot shows the Power BI desktop interface with a table view and a query settings pane.

Table View:

	ABC 123 director_name	ABC 123 num_critic_for_reviews	ABC 123 duration	ABC 123 director_facebook_likes	ABC 123 actor_3_facebook_likes
1	James Cameron	723	178	0	
2	Gore Verbinski	302	169	563	
3	Sam Mendes	602	148	0	
4	Christopher Nolan	813	164	22000	
5	Doug Walker	null	null	131	
6	Andrew Stanton	462	132	475	
7	Sam Raimi	392	156	0	
8	Nathan Greno	324	100	15	
9	Joss Whedon	635	141	0	
10	David Yates	375	153	282	
11	Zack Snyder	673	183	0	
12	Bryan Singer	434	169	0	
13	Marc Forster	403	106	395	
14	Andrew Adamson	258	150	80	
15	Rob Marshall	448	136	252	
16	Barry Sonnenfeld	451	106	188	
17	Peter Jackson	422	164	0	
18	Marc Webb	599	153	464	
19	Ridley Scott	343	156	0	
20	Chris Weitz	251	113	129	

Query Settings:

```
= Table.Distinct(#"Removed Columns", {"director_name"})
```

Properties:

- Name: Table1
- All Properties

Applied Steps:

- Source
- Removed Columns
- Removed Duplicates

Removal of rows with null values

Query Settings

Properties

- Name: Table1
- All Properties

Applied Steps

- Source
- Removed Columns
- Removed Duplicates
- Filtered Rows

Table View

Table Structure:

	get	title_year	actor_2_facebook_likes	imdb_score	movie_facebook_likes
1	237000000	2009	936	7.9	330
2	300000000	2007	5000	7.1	
3	245000000	2015	393	6.8	850
4	250000000	2012	23000	8.5	1640
5	263700000	2012	632	6.6	240
6	258000000	2007	11000	6.2	
7	260000000	2010	553	7.8	290
8	250000000	2015	21000	7.5	1180
9	250000000	2009	11000	7.5	100
10	250000000	2016	4000	6.9	1970
11	209000000	2006	10000	6.1	
12	200000000	2008	412	6.7	
13	225000000	2008	216	6.6	
14	250000000	2011	11000	6.7	580
15	225000000	2012	816	6.8	400
16	250000000	2014	972	7.5	650
17	230000000	2012	10000	7	560
18	200000000	2010	882	6.7	170
19					

Filter Row Count: 19

Split columns by delimiter

Queries [1]    = Table.SelectRows(#"Changed Type", each true)

Table1

	ABC 123 gross	ABC genres.1	ABC genres.2	ABC 123 actor_1_name
00%	Valid 100%	Valid 100%	Valid 95%	Valid 100%
0%	Error 0%	Error 0%	Error 0%	Error 0%
0%	Empty 0%	Empty 0%	Empty 5%	Empty 0%
1	1000	760505847 Action	Adventure Fantasy Sci-Fi	CCH Pounder
2	10000	309404152 Action	Adventure Fantasy	Johnny Depp
3	11000	200074175 Action	Adventure Thriller	Christoph Waltz
4	27000	448130642 Action	Thriller	Tom Hardy
5	640	73058679 Action	Adventure Sci-Fi	Daryl Sabara
6	24000	336530303 Action	Adventure Romance	J.K. Simmons
7	799	200807262 Adventure	Animation Comedy Family Fantasy Musical Romantic	Brad Garrett
8	26000	458991599 Action	Adventure Sci-Fi	Chris Hemsworth
9	25000	301956980 Adventure	Family Fantasy Mystery	Alan Rickman
10	15000	330249062 Action	Adventure Sci-Fi	Henry Cavill
11	18000	200069408 Action	Adventure Sci-Fi	Kevin Spacey
12	451	168368427 Action	Adventure	Giancarlo Giannini
13	22000	141614023 Action	Adventure Family Fantasy	Peter Dinklage
14	40000	241063875 Action	Adventure Fantasy	Johnny Depp
15	10000	179020854 Action	Adventure Comedy Family Fantasy Sci-Fi	Will Smith
16	5000	255108370 Adventure	Fantasy	Aidan Turner
17	15000	262030663 Action	Adventure Fantasy	Emma Stone
18	891	105219735 Action	Adventure Drama History	Mark Addy
19				

21 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 112

Query Settings

PROPERTIES

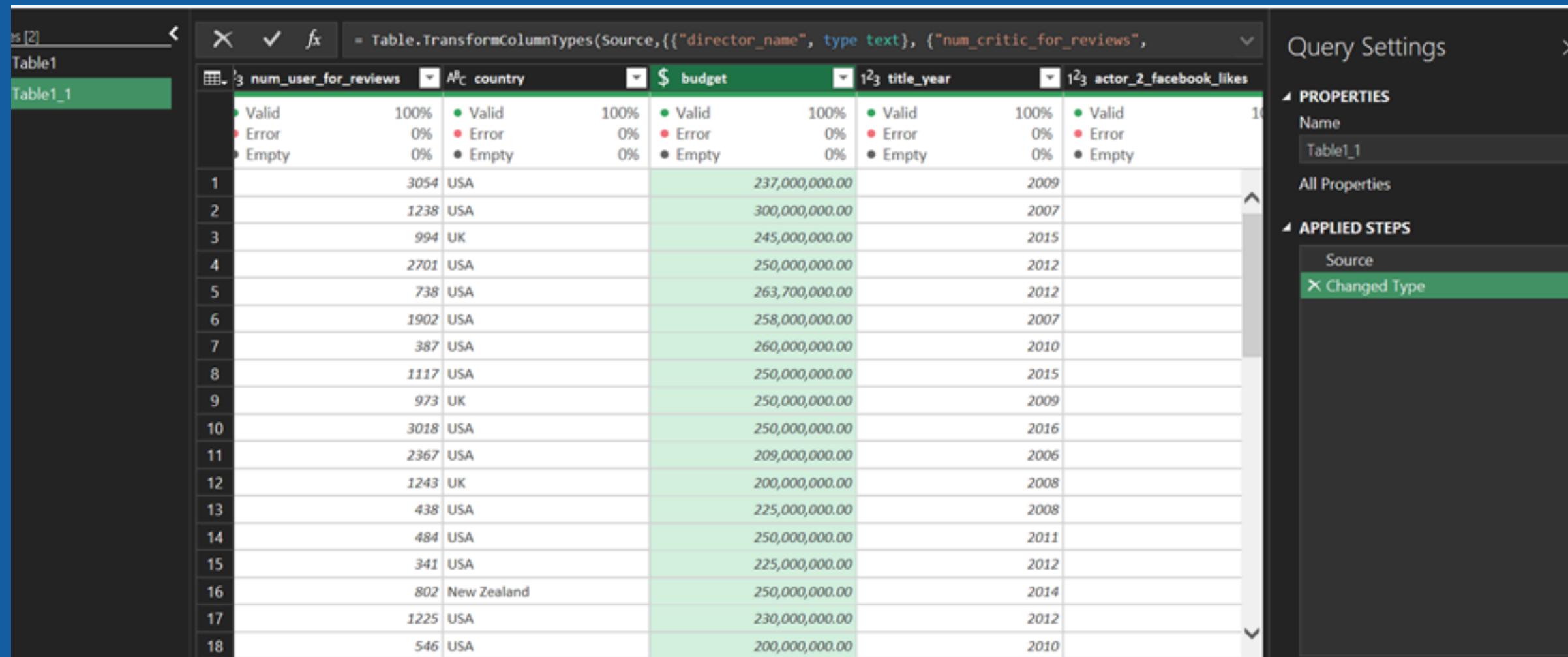
Name: Table1

All Properties

APPLIED STEPS

- Source
- Removed Columns
- Removed Duplicates
- Filtered Rows
- Replaced Value
- Split Column by Delimiter
- Changed Type
- Filtered Rows1

Change data types



The screenshot shows the Power BI Query Editor interface. On the left, there is a table named 'Table1_1' with 18 rows of data. The columns are: 'num_user_for_reviews', 'country', '\$ budget', 'title_year', and 'actor_2_facebook_likes'. The 'country' column contains values like 'USA', 'UK', and 'New Zealand'. The '\$ budget' column contains values like '237,000,000.00' and '200,000,000.00'. The 'title_year' column contains years from 2006 to 2014. The 'actor_2_facebook_likes' column contains values like '546' and '1225'. Above the table, the formula bar shows: `= Table.TransformColumnTypes(Source,{{"director_name", type text}, {"num_critic_for_reviews",`. To the right of the table is the 'Query Settings' pane, which includes sections for 'PROPERTIES' (Name: Table1_1) and 'APPLIED STEPS' (Source, Changed Type). The 'Changed Type' step is highlighted with a green background.

DESCRIPTIVE ANALYSIS

In this section we look at the mean, median and range values of Budget, Gross, IDMB Score, Duration and Movie Facebook Likes

Descriptive Analysis		
Particulars	Average	Median
Budget	181623232.3	175000000
Gross	192773425.9	161087183
IDMB Score	6.753535354	6.7
Duration	122.7474747	119
Movie Facebook Likes	36040.40404	24000

Particulars	Range	Maximum	(Minimum)	Range_Value
Budget	300000000	100000000	200000000	200000000
Gross	760505847	21379315	739126532	739126532
IDMB Score	8.5	4.2	4.3	4.3
Duration	206	85	124	124
Movie Facebook Likes	197000	0	197000	197000

Insights from the descriptive analysis

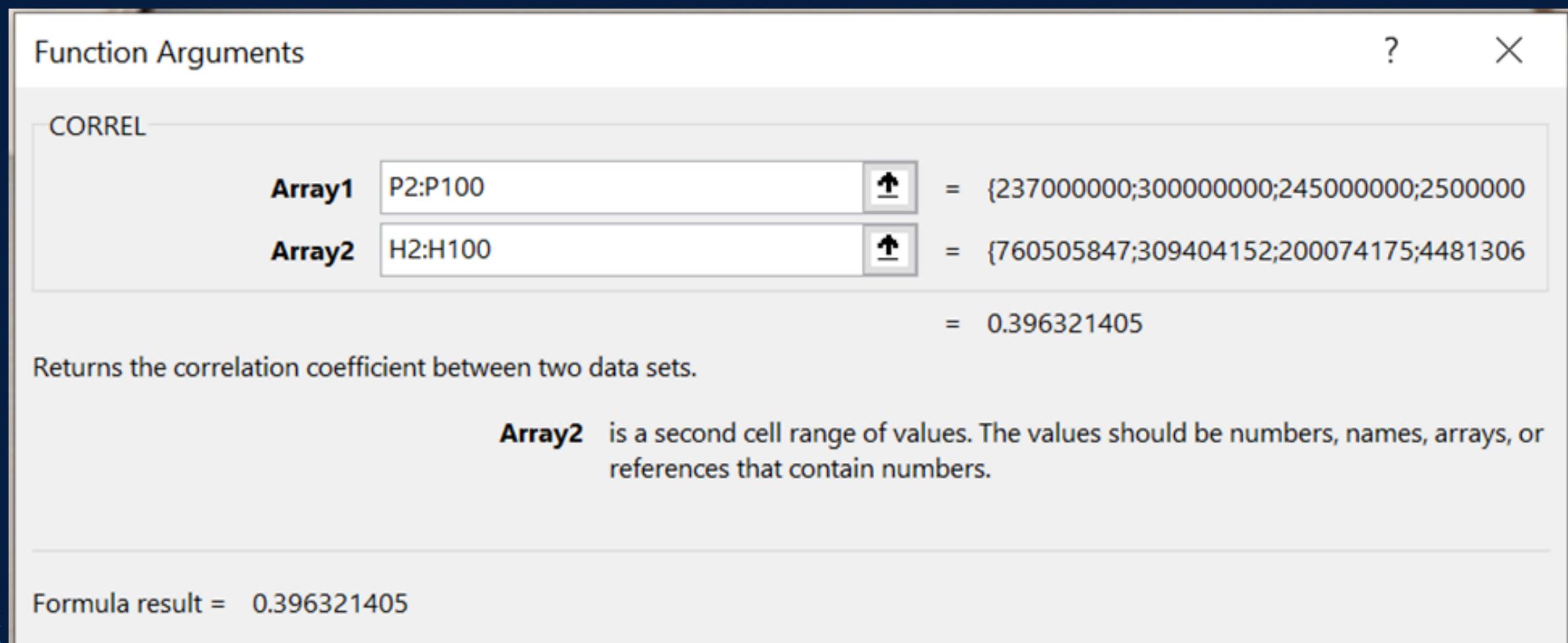
Budget: The average budget is around \$181.62 million, with a minimum of \$100 million and a maximum of \$300 million.

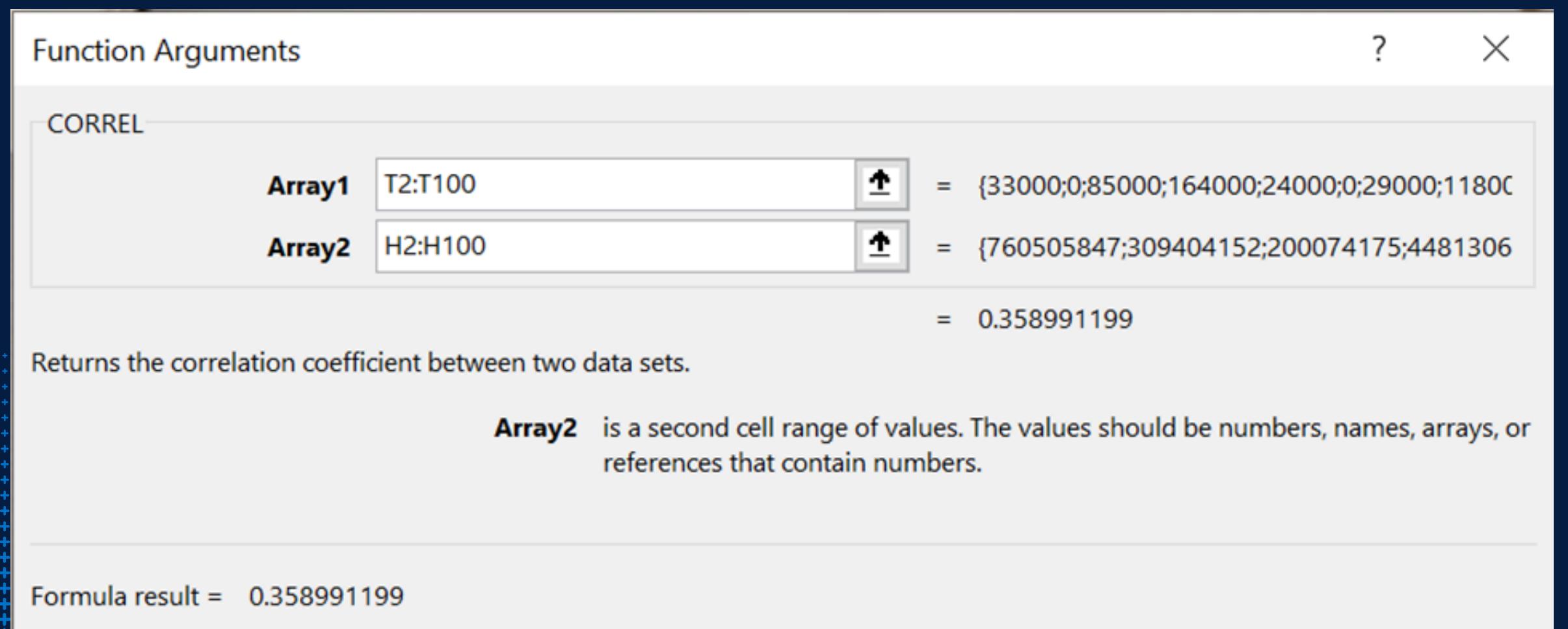
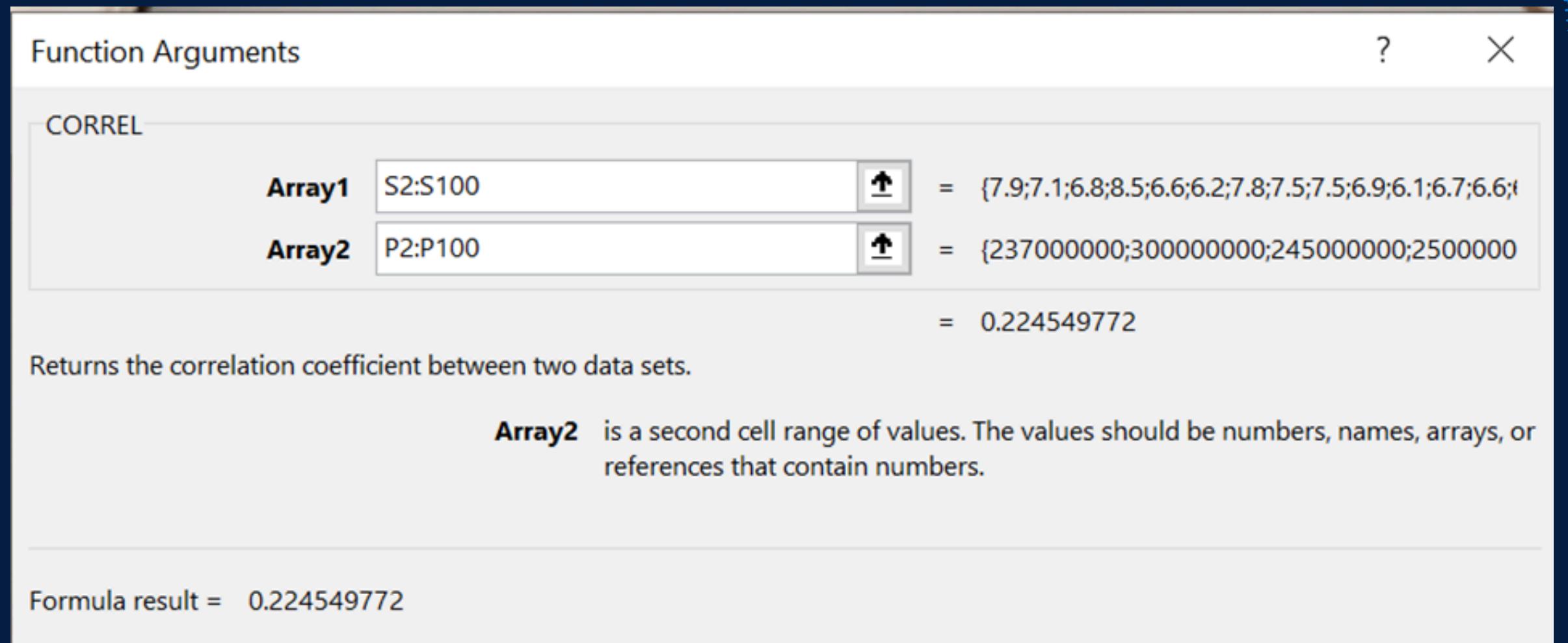
Gross Earnings: The average gross revenue is \$192.77 million, with a maximum of \$760.51 million and a minimum of \$213.79 million.

IMDB Score: The average IMDB score is 4.3, ranging from 4.2 to 8.5.

CORRELATION ANALYSIS

Here, we look at the relationships that exist between Budget Vs Gross, IDMB Score Vs Budget and Movie Facebook Likes Vs Gross





Insights from the correlation analysis

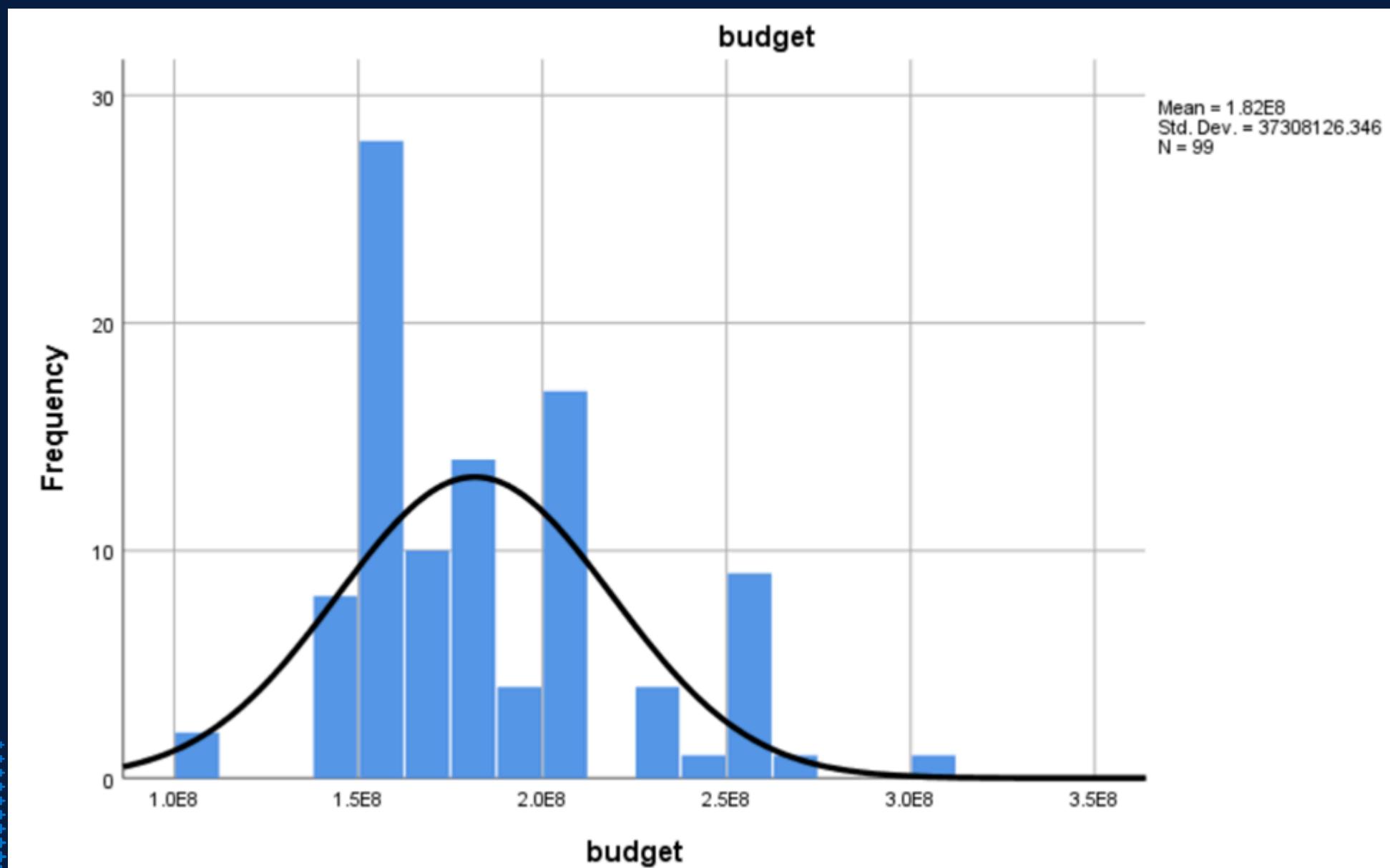
Budget vs Gross: There is a positive correlation (0.40), meaning higher budgets does guarantee higher gross earnings.

IMDB Score vs Gross: Moderate positive correlation (0.22), indicating that higher-rated movies tend to perform better at the box office.

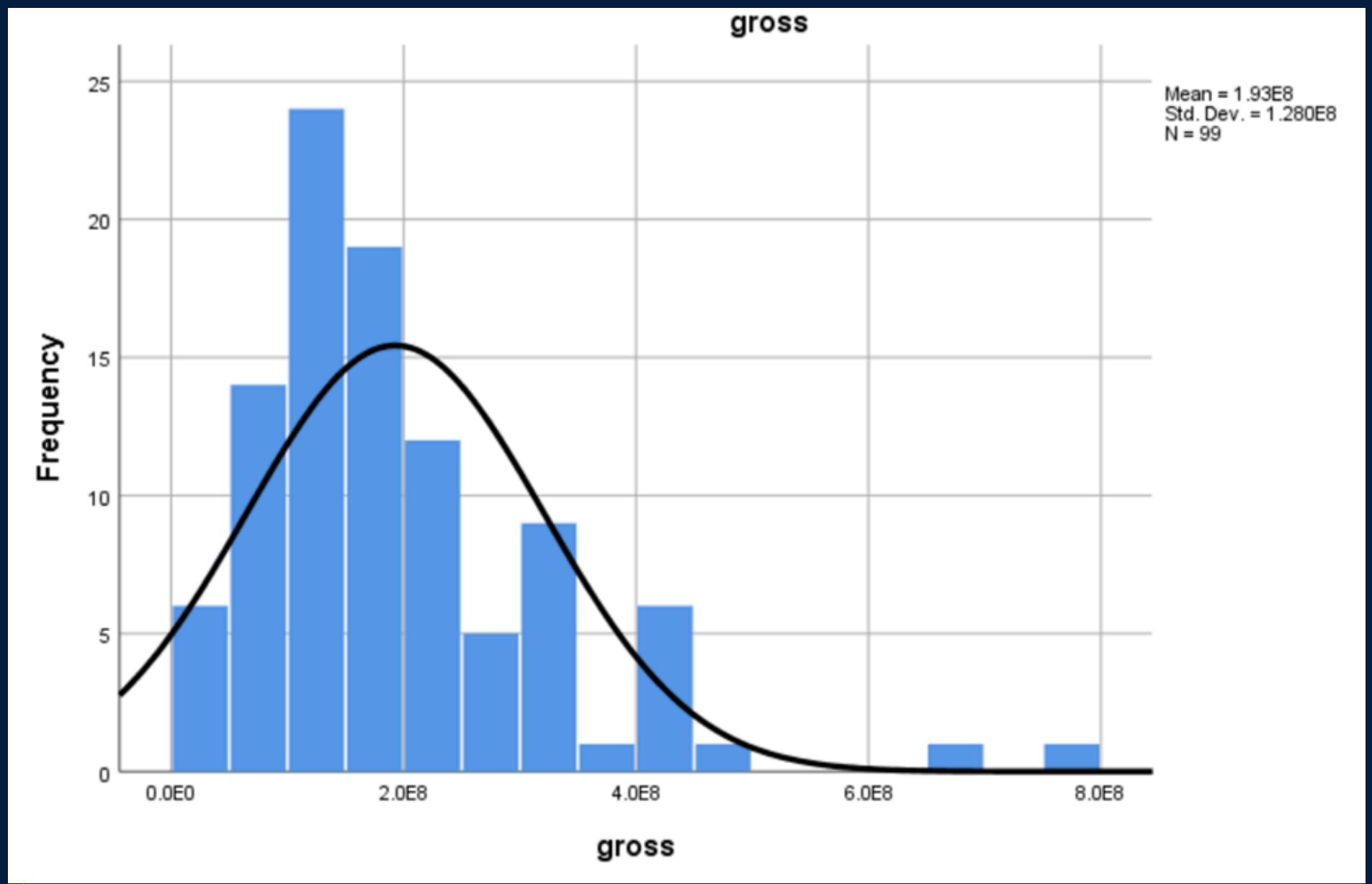
Social Media Impact: There is a strong positive correlation between gross earnings and movie Facebook likes (0.40) suggesting that movies with larger social media presences tend to be associated with higher-box office performances.

VISUALIZATIONS

Histograms of Budget



Histograms of Gross earnings

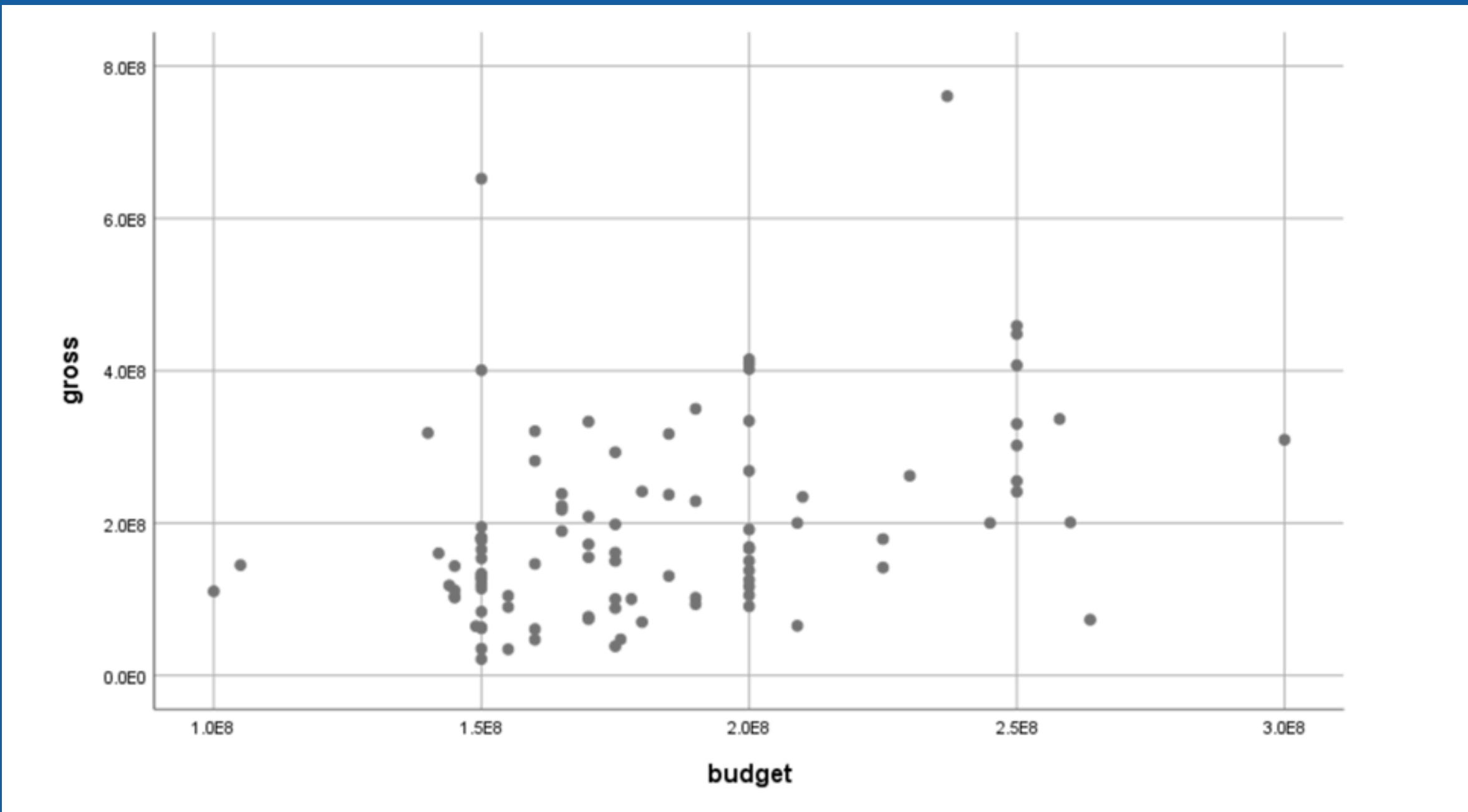


Insights from the histograms

The distribution of budgets shows that all the movies analyzed are mixed of moderate to high-budget blockbusters.

Gross earnings have a wide range, with a few movies earning significantly more than others.

Scatter Plots of Budget Vs. Gross



Insights from scatter plot

The scatter plot shows that while budget and gross earnings do not follow a direct linear trend, some high-budget movies did achieve high gross revenue.

Heat Map

	<i>duration</i>	<i>budget</i>	<i>gross</i>	<i>mdb_score</i>	<i>facebook_likes</i>
<i>duration</i>	1				
<i>budget</i>	0.465431	1			
<i>gross</i>	0.293289	0.396321	1		
<i>imdb_score</i>	0.114431	0.22455	0.465188	1	
<i>movie_facebook_likes</i>	0.29984	0.159392	0.358991	0.474635	1

Insights from heat map

The heatmap highlights the relationships among duration of movies, budget, gross earnings, IMDB scores, and social media presence. It confirms that social media metrics for movies correlate strongly with IMDB scores.

RECOMMENDATION

From the various insights uncovered after an indepth exploration of data analysis reveals that in as much higher budgets for movies are good, they do not necessarily guarantee the success of movies at the box-office.

Movie owners need to engage high performing players ranging from directors to actors in order to ensure the success of their movies.

Also there is need for effective PR most especially on social media platforms to engage netizens as more movie lovers tend to be on the various social media platforms. It has been proven that social media does drive movie success.