Name: Nikhil Zodape

Email Id.: Nikszodape@gmail.com

Project Name: IMDB Movie Analysis

Using: Python

Problem Statement:

The dataset at hand pertains to IMDB Movies, prompting the exploration of a key question: "What factors contribute to the success of a movie on IMDB?" In this context, success is gauged by elevated IMDB ratings. The implications of unravelling this issue are substantial for stakeholders in the film industry, such as producers, directors, and investors. A deeper understanding of the elements influencing a movie's success is crucial for making informed decisions in future projects.

Procedure

Data Cleaning:

One of the most important step to perform before moving forward with the analysis.

➔ Removing irrelevant data:

1. This task can be performed by reading the project description and by understanding which columns are required in our further analysis. Once we get to know which columns are important, we will remove all the unnecessary columns from the dataset.
2. There are total 28 columns present in the dataset after removing 18 columns we are having 10 columns to work with.
3. The 18 removed columns are: Color, num_critic_for_reviews, actor_2_name, director_facebook_likes, actor_3_facebook_likes, actor_1_facebook_likes, cast_total_facebook_likes, facenumber_in_poster, plot_keywords, movie_imdb_link, actor_2_facebook_likes, movie_facebook_likes, actor_3_name, num_voted_users, num_user_for_reviews, content_rating, aspect_ratio

➔ Removed Duplicated:

1. We removed the duplicated rows which was 122.
2. Removed duplicated movies title rows which was 4.

➔ Checking NaN values in the dataset.

```
director_name     104
duration           15
gross             884
genres              0
actor_1_name        7
movie_title         0
language           12
country             5
budget            492
imdb_score          0
dtype: int64
```

```
gross             17.529248
budget             9.756098
director_name      2.062265
duration           0.297442
language           0.237954
actor_1_name       0.138806
country            0.099147
genres             0.000000
movie_title        0.000000
imdb_score         0.000000
dtype: float64
```

We will examine the NaN values in the form of percentages to facilitate additional analysis.

```
Steven Spielberg    26
Woody Allen         22
Clint Eastwood      20
Martin Scorsese     20
Ridley Scott        16
Spike Lee           16
Steven Soderbergh   15
Renny Harlin        15
Oliver Stone        14
Tim Burton          14
Ron Howard          13
Name: director_name, dtype: int64
```

Handling Missing Data:

➔ **Director Names**

1. The "Director" column contains NaN values, and leaving the director name column blank is not viable for our subsequent analysis, as it is essential information. Given that this column represents a categorical variable, filling it with the mean is not appropriate. Instead, we will employ the value count method/mode. We plan to address this by selecting the top 11 directors from the dataset and then filling in the blank entries in the 102 rows with the names of these directors

.

➔ **Language**

2. We have addressed the language column by populating it based on the count of top languages in movies. All 12 NaN values in the language column have been filled with "English."

```
language
English      4586
French         73
Spanish        40
Hindi          28
Mandarin       24
German         19
Japanese       17
Italian        11
Cantonese      11
Russian        11
Name: language, dtype: int64
```

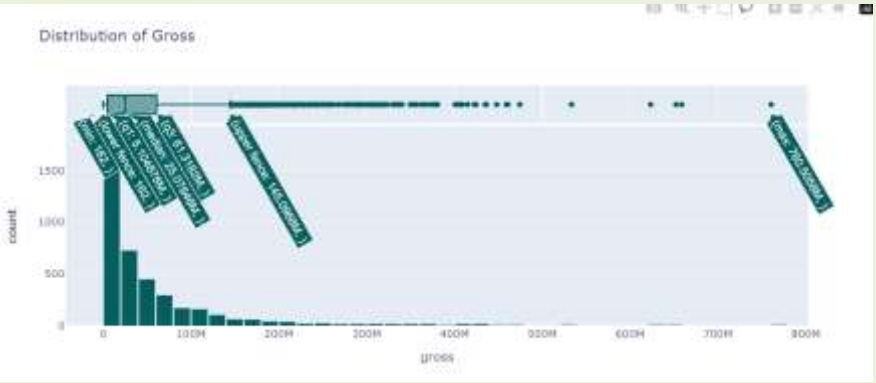
```
gross              17.577728
budget              9.855720
duration            0.304816
actor_1_name        0.142248
country             0.101605
director_name       0.000000
genres              0.000000
movie_title         0.000000
language            0.000000
imdb_score          0.000000
dtype: float64
```

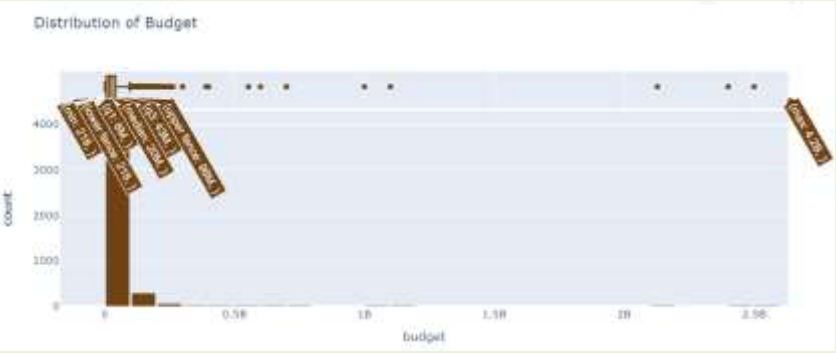
➔ **Dropping NaN values which are less than 1%.**

3. We are now proceeding to eliminate all NaN values that account for less than 1% of the total values in the dataset.

➔ **Filling NaN Values of Numerical Columns i.e. Gross and Budget**.

V

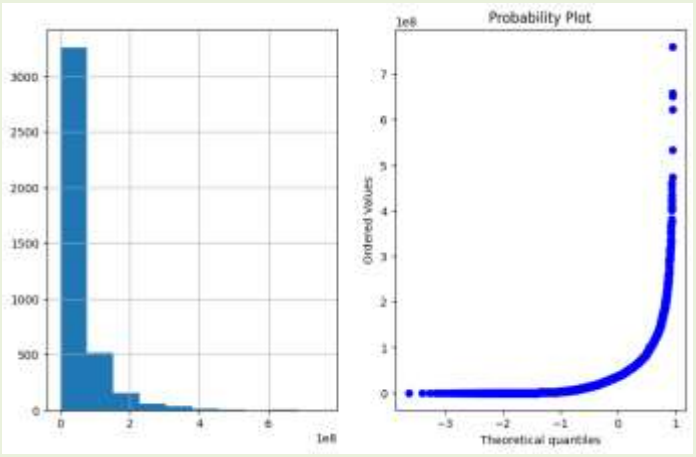




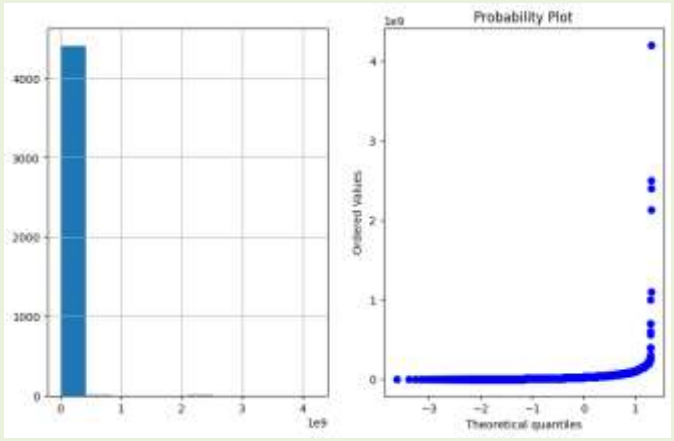
4. Prior to addressing the NaN values in the Gross and Budget columns, we initiated the process by generating Box and distribution plots to visualize the data distribution. Upon observing that the data is Right Skewed, we opted for Log transformation to achieve a Gaussian distribution. Subsequently, we filled the NaN values using the Median, ensuring a well-balanced representation of our data.

➔ **Distribution of data before Log-Transformation using Q-Q Plot**

Gross

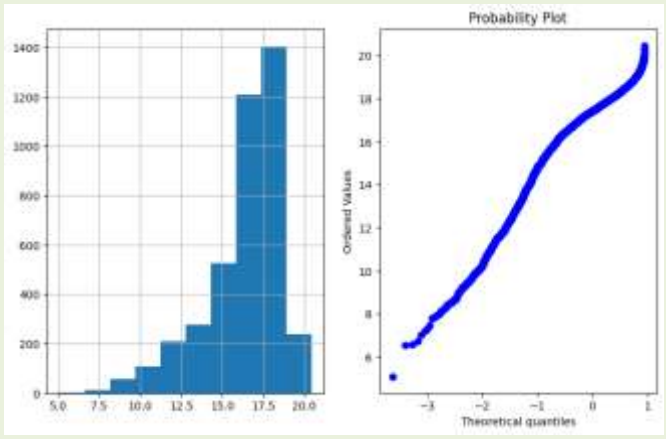


Budget

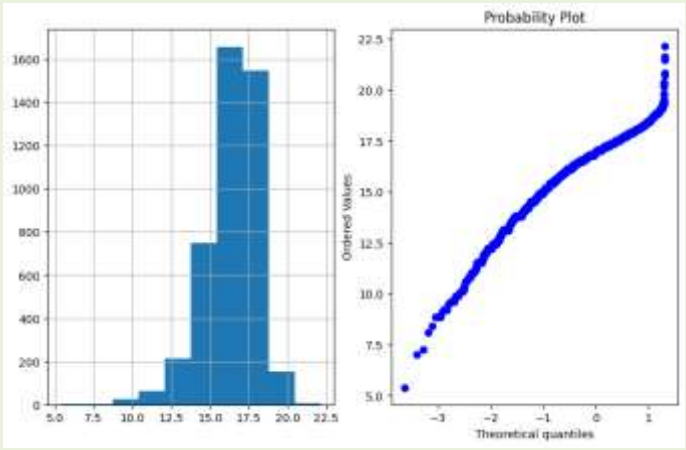


➔ **Distribution of data After Log-Transformation using Q-Q Plot**

Gross



Budget



➔ **Data Cleaning and filling NaN values of Genres Column.**

5. In order to enhance the analysis of the Genre column, which was initially filled using pipes, we performed a separation of these pipes. This resulted in the creation of individual columns for each genre, namely genre_1, genre_2, genre_3, genre_4, genre_5, genre_6, genre_7, and genre_8, as illustrated in the figure below.

	genre_1	genre_2	genre_3	genre_4	genre_5	genre_6	genre_7	genre_8
0	Action	Adventure	Fantasy	Sci-Fi	None	None	None	None
1	Action	Adventure	Fantasy	None	None	None	None	None
2	Action	Adventure	Thriller	None	None	None	None	None
3	Action	Thriller	None	None	None	None	None	None
5	Action	Adventure	Sci-Fi	None	None	None	None	None
...	...	...	...	...	...	...	...	...
5038	Comedy	Drama	None	None	None	None	None	None
5039	Crime	Drama	Mystery	Thriller	None	None	None	None
5040	Drama	Horror	Thriller	None	None	None	None	None
5041	Comedy	Drama	Romance	None	None	None	None	None
5042	Documentary	None	None	None	None	None	None	None

4891 rows × 8 columns

6. We are currently inspecting all NaN values across the various columns, and we will address them systematically based on our findings.
7. Following our examination, it has come to our attention that numerous NaN values exist in columns ranging from genre_3 to genre_8. For our subsequent analysis, we have decided to focus on genre_1 and genre_2 columns. To address the NaN values in genre_2, we will be filling them with the corresponding values from the genre_1 column.

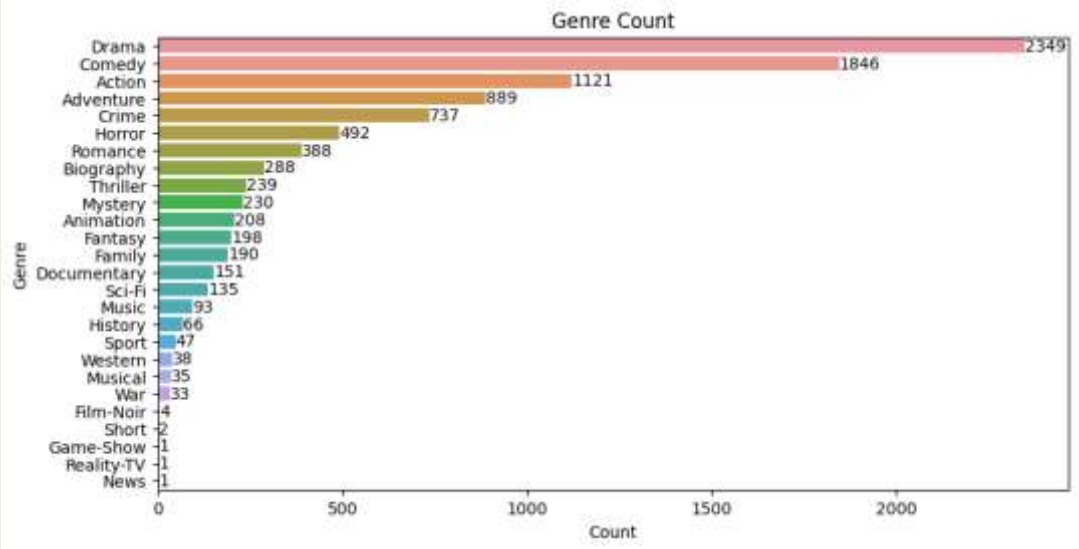
```
genre_1       0
genre_2     612
genre_3    1932
genre_4    3506
genre_5    4454
genre_6    4796
genre_7    4869
genre_8    4887
dtype: int64
```

Having completed the data cleaning process, we are now prepared to proceed with the subsequent Data Analysis tasks.

A. Movie Genre Analysis:

Analyze the distribution of movie genres and their impact on the IMDB score.

- Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores



The most common genres of moves is Drama, Comedy, Action, Adventure, Crime, Horror and Romance of it IMDB Score.

		count	mean	std	min	25%	50%	75%	max
genre_1	genre_2								
Comedy	Drama	513.0	6.557115	0.836796	3.3	6.000	6.70	7.200	8.8
Action	Adventure	448.0	6.345759	1.136374	1.9	5.800	6.40	7.000	8.9
Crime	Drama	283.0	6.991519	0.931900	3.4	6.400	7.10	7.600	9.3
Drama	Drama	233.0	6.911588	0.922045	3.3	6.500	7.10	7.500	9.1
Action	Crime	205.0	6.283415	0.923027	2.8	5.900	6.40	6.800	9.0
Comedy	Comedy	204.0	5.827451	1.288307	1.9	5.200	5.90	6.625	9.5
Drama	Romance	196.0	6.938776	0.789889	3.5	6.500	7.00	7.500	8.6
Biography	Drama	175.0	7.172000	0.646266	5.3	6.800	7.20	7.600	8.9
Comedy	Romance	171.0	5.888889	0.944859	2.7	5.400	5.90	6.500	8.4
Action	Comedy	150.0	5.939333	0.914895	3.5	5.325	6.00	6.600	7.9
Comedy	Crime	149.0	6.259060	1.042672	2.4	5.800	6.30	6.800	8.4
Adventure	Animation	133.0	6.543609	1.112011	2.8	5.900	6.70	7.300	8.6
	Comedy	115.0	6.207826	1.098016	2.3	5.450	6.20	7.000	8.5
Action	Drama	109.0	6.438532	1.088994	3.1	5.800	6.50	7.100	8.6
Adventure	Drama	84.0	6.957143	1.006639	4.0	6.300	7.20	7.700	8.6
Drama	Mystery	83.0	6.828916	0.923478	4.4	6.200	6.90	7.500	8.5
Comedy	Family	82.0	5.600000	1.255900	1.9	4.950	5.75	6.275	8.7
Horror	Mystery	71.0	5.850704	1.004827	3.9	5.150	5.80	6.450	8.5
	Horror	67.0	5.501493	1.230237	2.2	4.650	5.60	6.350	8.0
Drama	Thriller	64.0	6.435937	1.011892	3.9	6.000	6.55	7.000	8.5

B. Movie Duration Analysis:

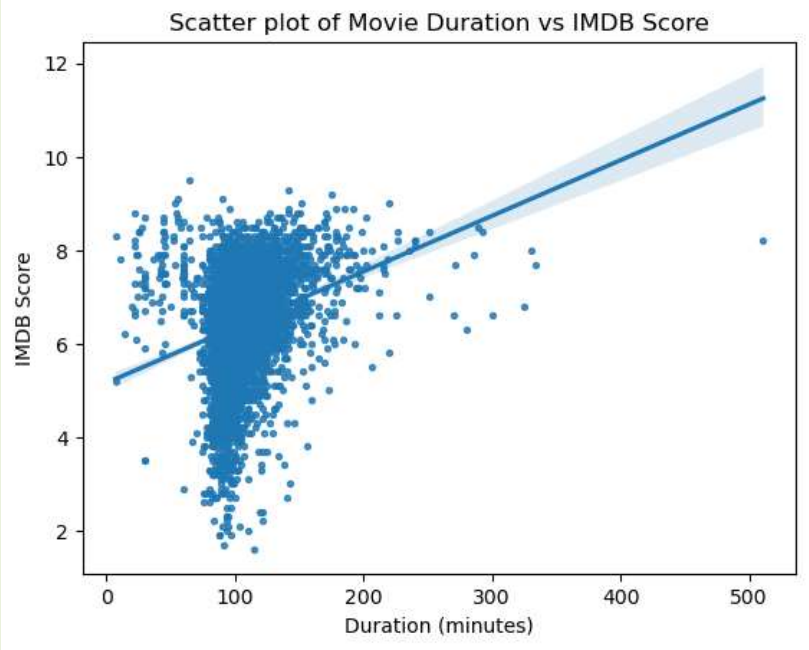
Analyze the distribution of movie duration and its impact on the IMDB score.

- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB Score.

```
mean      107.154979
median    103.000000
std        25.242523
Name: duration, dtype: float64
```

The average movie duration is 107 minutes.

The duration of movie data is normally distributed since the mean and median are close to each other's.



The trendline is scatterplot has slightly upward but not to much and it indicates that it has moderate positive relationship between duration and imdb score.

C. Language Analysis:

Situation: Examine the distribution of movies based on their language.

- Task: Determine the most common languages used in movies and analyze their impact on the IMDB Score using descriptive statistics.

language	Count	Mean	Median	Std
English	4575	6.394973	6.50	1.13
French	73	7.038356	7.20	0.72
Spanish	40	6.937500	7.15	0.84
Hindi	27	6.774074	7.00	1.18
Mandarin	24	6.787500	7.05	1.02
German	19	7.342105	7.60	0.93
Japanese	17	7.347059	7.50	0.97
Russian	11	6.363636	6.50	1.32
Cantonese	11	6.954545	7.20	0.67
Italian	11	7.227273	7.30	1.19

As we can see than English is the most common language used in 4575 movies, followed by French with 73 movies and Spanish with 40 movies.

Telugu, Polish, Indonesian languages have the highest mean of 8.4, 7.96 and 7.9.

language	Count	Mean	Median	Std
Telugu	1	8.400000	8.40	0.00
Polish	3	7.966667	7.40	0.80
Indonesian	2	7.900000	7.90	0.30
Maya	1	7.800000	7.80	0.00
Hebrew	5	7.580000	7.60	0.30

D. Director Analysis:

Influence of directors on movie ratings.

- Task: Identify the top directors based on their average IMDB Score and analyze their contribution to the success of movies using percentile calculation.

```
The 75% percentile for overall distribution of IMDB score with respect to director name is 7.2

There are 461 director who have the average of 7.59 who are in the 75 percentile
```

```
director_name
John Blanchard          9.500
Cary Bell               8.700
Mitchell Altieri        8.700
Sadyk Sher-Niyaz        8.700
Charles Chaplin         8.600
Mike Mayhall            8.600
Damien Chazelle         8.500
Majid Majidi            8.500
Raja Menon              8.500
Ron Fricke              8.500
Sergio Leone            8.475
Tony Kaye               8.450
Christopher Nolan       8.425
Asghar Farhadi          8.400
Bill Melendez           8.400
Catherine Owens         8.400
Jay Oliva               8.400
Marius A. Markevicius   8.400
Moustapha Akkad         8.400
Rakeysh Omprakash Mehra 8.400
Name: imdb_score, dtype:  float64
```

From the percentile we can say that any values greater than 7.2 Imdb socre will be considered as the top director.

E. Budget Analysis:

Explore the relationship between movie budget and their financial success.

- Task: Analyse the correlation between movie budget and gross earnings, and identify the movies with the highest profit margin.

```
Correlation coefficient between budget and gross: 0.239.

Note: All the values of profit margin are in the `Million $`.
```

	movie_title	profit_margin
0	Avatar	523.505847
29	Jurassic World	502.177271
26	Titanic	458.672302
3024	Star Wars: Episode IV - A New Hope	449.935665
3080	E.T. the Extra-Terrestrial	424.449459
17	The Avengers	403.279547
509	The Lion King	377.783777
240	Star Wars: Episode I - The Phantom Menace	359.544677
66	The Dark Knight	348.316061
439	The Hunger Games	329.999255

Link of complete Python workbook: https://drive.google.com/file/d/1ePd5QXMn0eajuIQa-AfqSaRI-fdAILzt/view?usp=drive_link

Video Link: https://drive.google.com/file/d/1iBLFnqIiCMjzusX9MxyCnpHjDN38IanV/view?usp=drive_link