ANALYSIS OF TMDB DATASET

The tmdb dataset is a dataset that contains 108066 rows and 21 columns, it describes various movies and details about them including: casts, ratings, movie director, budget, the revenue etc. This project aims to analyse the dataset and draw some meaningful insights. To do this, 4 main steps were carried out:

1. **STEP ONE: ASKING QUESTIONS**
   Below are the questions looked at in order to draw meaningful insights from the data.

```
Q1. How much has the rating improved over the years

Q2. Visualize the popularity of the movies over the years

Q3. What level of popularity receives the highest revenue?

Q4. What level of runtime associated with the highest voting score?

Q5. Who are the top 5 cast memebers with highest movie voting (check
 only first 200 top voted movies)?

Q6. Which genres are most popular

Q7: What are the top 5 movies with the highest profit
```

2. **STEP TWO: DATA WRANGLING -INSPECTING THE DATA**

In this section we will do the following:

1. Drop unnecessary columns: We dropped homepage, tagline, overview, id, imdb_id, vote_count, and keywords, because these columns do not add any meaningful insight to our data.
2. Dealing with missing data: Some rows with missing data were dropped while others were filled with data. A column like cast is difficult to fill, so rows with empty cast values were dropped. Furthermore, upon analysis of rows with missing "director" values, it summed up to 38 rows (which is the second highest frequency after "Woody Allen") as seen in the figure 1 below. So instead of looking such data that may be important. I assigned missing director rows to a new category called "Missing". Next, for the genre, "Comedy" and "Drama" have the highest frequencies, with Comedy surpassing Drama by just one. Since we have 22 rows with missing genres, I assigned 11 missing rows to the genre category "Comedy" and the other 11 to "Drama".
3. Check for duplicates: Just one duplicate column was found, and this column was dropped.
4. Ensure correct datatype of the column: release_year was the only column with a wrong datatype. The datatype was changed to the correct datetime format.

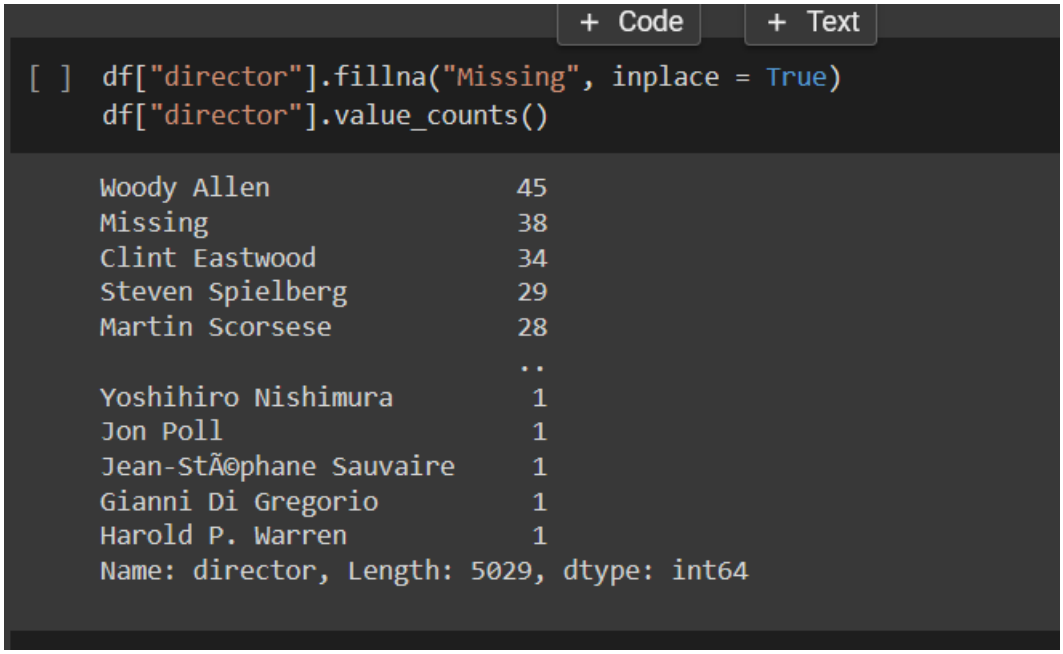


Figure 1:Dealing with null values on "Director" column

3. STEP THREE: EXPLORATORY DATA ANALYSIS(EDA)

In this section, we try to understand the data and relationships that exist between columns so we can draw insights and answer the questions posed.

a. First, a histogram was plotted for all numerical data. From the histogram plots below, we can see that most of the features are not normally distributed. They are mostly skewed to the left with an exception of vote average and release year.

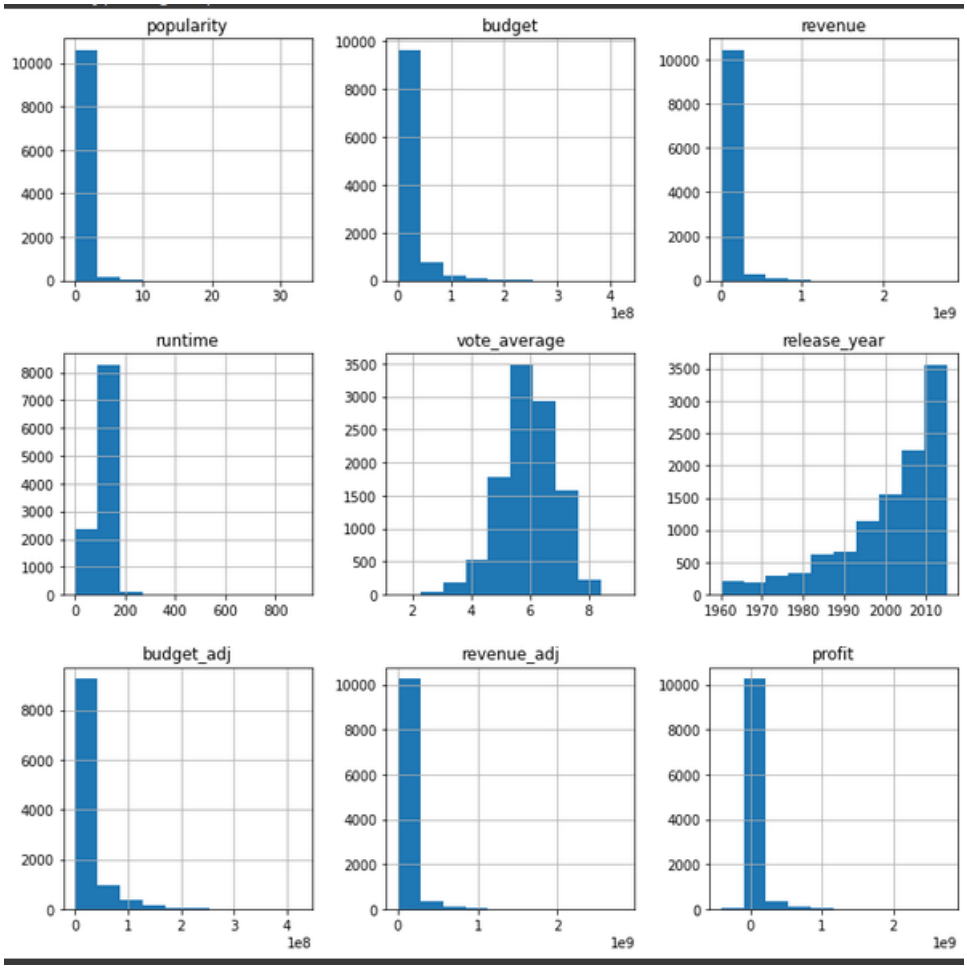


Figure 2: Histogram plots

b. Next, we try to analyze the correlation between some features and we see that there is no much correlation between the numerical features. An example is correlation between budget and profit(lets try to see if the more budget allocated to a movie, the more profit the movie get or vice versa). From the graph below, we can see that there is no positive or negative correlation between these two variables so we cant come to a concrete conclusion. Same goes to other features such as release year vs popularity, and so on.

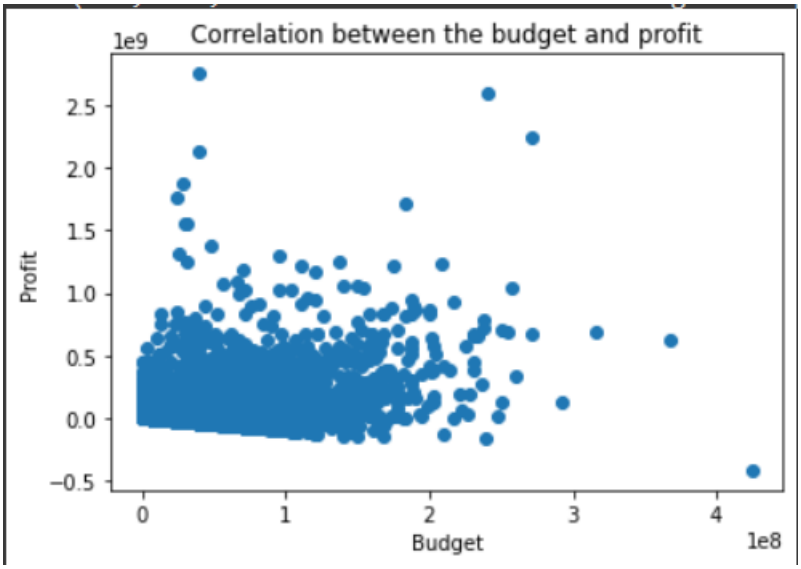
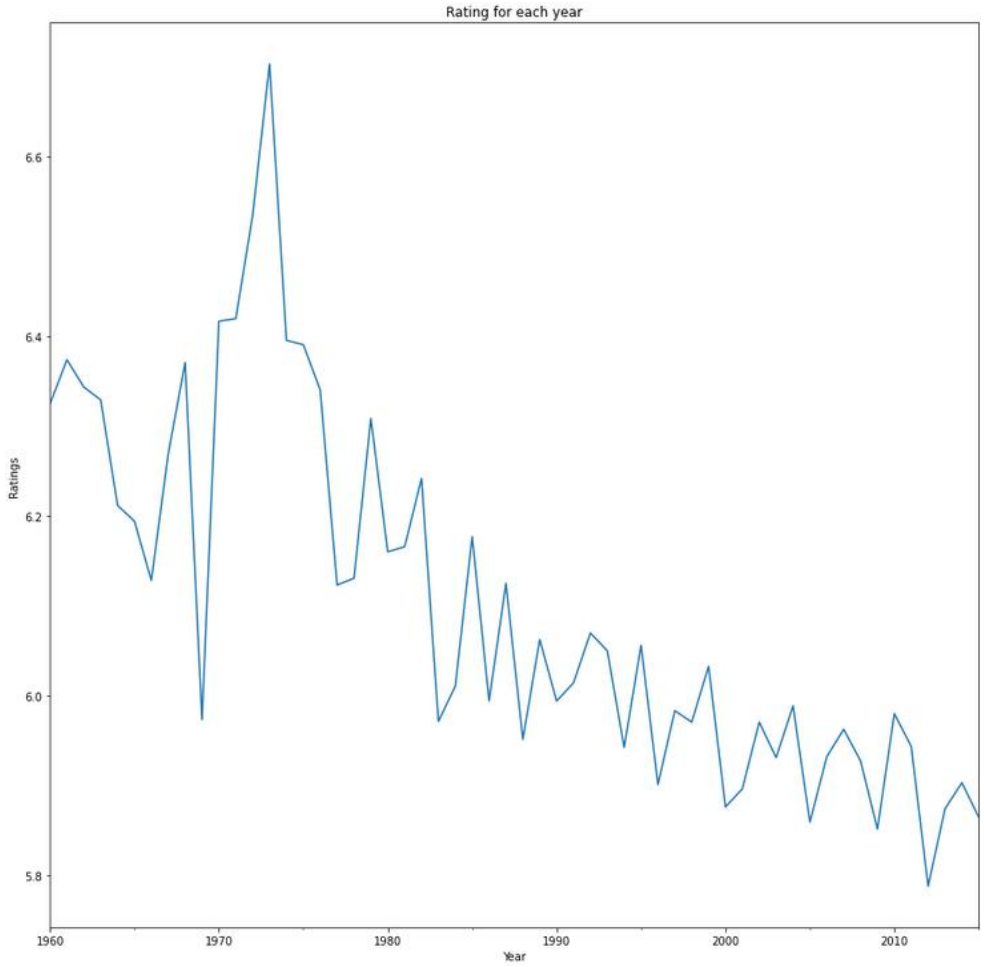


Figure 3: Correlation between budget and profit.

c. In the final stage of EDA, we do some calculations and plot some graphs to answer the questions posed.

Q1. How much has the rating improved over the years

```
[ ]  # We need to calculate the overall rating for each year
     df.groupby("release_year")["vote_average"].mean().plot( kind = "line", figsize = (15,15));
     plt.xlabel("Year")
     plt.ylabel("Ratings")
     plt.title("Rating for each year")
```



Similar computation was done for the other six questions(can be seen in the attached ipynb file).

4. STEP FOUR: DRAWING CONCLUSIONS

In this section, we will come to a conclusion regarding the questions posed.

1. First, from the histogram, we notice that all the numerical features are not normally distributed. This can also be seen as a limitation in the dataset .
2. Below are the insights gotten from the data based on the questions posed and also the limitations

INSIGHTS

Q1. Between the year 1960 and 2015, there was an overall decrease in the average ratings of all movies. with a value of 0.4

Q2. Over the years there is an increase in the popularity of the movies with about 0.58

Q3. Movies with higher popularity had higher revenue

Q4. The top 5 cast members with highest movie voting are: Elijah wood, Micheal Caine, Jared Leto, ian Mckellen, Robert De Niro

Q5. The 5 most popular genres are: Drama, Documentary, Music Comedy and crime

Q6. The five movies with the highest profit are: Star wars, Avatar, Titanic, The exorsist and Jaws.

LIMITATIONS

1. There is no much correlation between the numerical features in the dataset

2. Since most features are categorical, we cannot come to a conclusion regarding the correlation between such features