PROJECT 1: INVESTIGATE A DATASET- TMDB **REPORT**

JUNE 2022 // PREPARED BY ENOCK CHESIRE

INTRODUCTION

For this project 1, I chose to work on the Tmdb dataset. It contains information about 10,000 movies collected from The Movie Database (Tmdb). It contains information such as revenue, popularity of movies, budget, cast, title, runtime, genres, and release date.

The dataset has 6621 rows and 21 columns as shown in the image below. .

```
RangeIndex: 6621 entries, 0 to 6620
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    6621 non-null   int64
 1   imdb_id               6613 non-null   object
 2   popularity            6621 non-null   float64
 3   budget                6621 non-null   int64
 4   revenue               6621 non-null   int64
 5   original_title        6621 non-null   object
 6   cast                  6565 non-null   object
 7   homepage              2423 non-null   object
 8   director              6583 non-null   object
 9   tagline               4678 non-null   object
 10  keywords              5465 non-null   object
 11  overview              6616 non-null   object
 12  runtime               6620 non-null   float64
 13  genres                6603 non-null   object
 14  production_companies  5915 non-null   object
 15  release_date          6620 non-null   object
 16  vote_count            6620 non-null   float64
 17  vote_average          6620 non-null   float64
 18  release_year          6620 non-null   float64
 19  budget_adj            6620 non-null   float64
 20  revenue_adj           6620 non-null   float64
dtypes: float64(7), int64(3), object(11)
```

Some columns such as cast and genres have multiple values. The homepage, overview and tagline columns have some missing values also.

QUESTIONS

The following were the guiding questions in my analysis:

1. What effect does runtime have on popularity?
2. What has been the trend in runtime over the years?
3. What is the trend in the number of movie productions over the years?
4. What is the trend in revenue over the years?
5. What is the relationship between revenue and popularity?
6. What is the relationship between revenue and budget?
7. Which genre has more releases?
8. What is the relationship between revenue, popularity, runtime, and budget?

FINDINGS

This section describes the findings of the Tmdb dataset that was being analysed

1. What is effect does runtime have on popularity?

From Figure 1, most movies with a runtime between 100 and 200 minutes have a higher popularity.

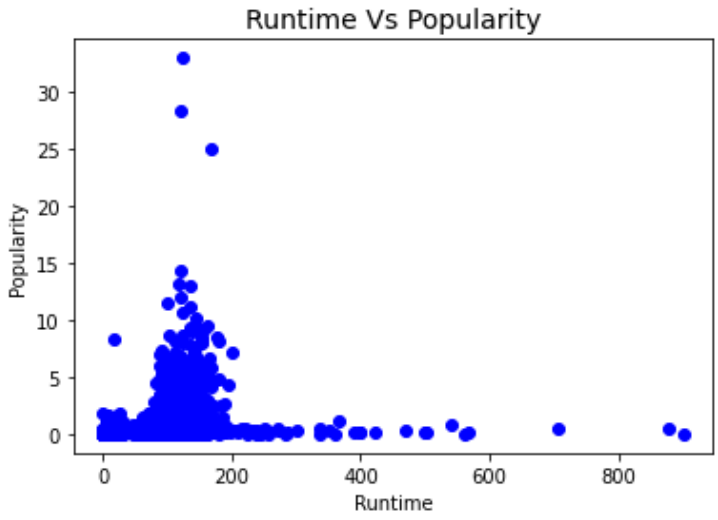

Figure 1.

2. What has been the trend in runtime over the years?

From the Figure 2, it looks like the runtime of movies being produced has been decreasing over the years. This could be due to movies having a runtime between 100 and 200 being more popular.

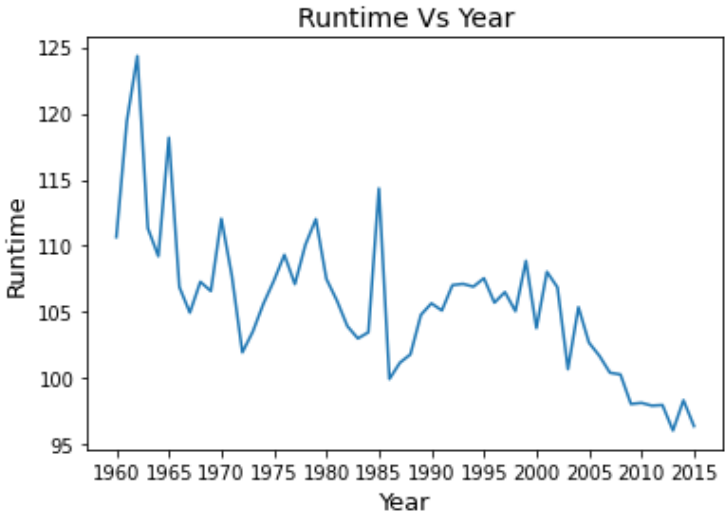

Figure 2.

3. What is the trend in the number movie production over the years?

There has been an increase in the number of movies being produced over the years most probably due to increase in production houses and the rise of streaming platforms.

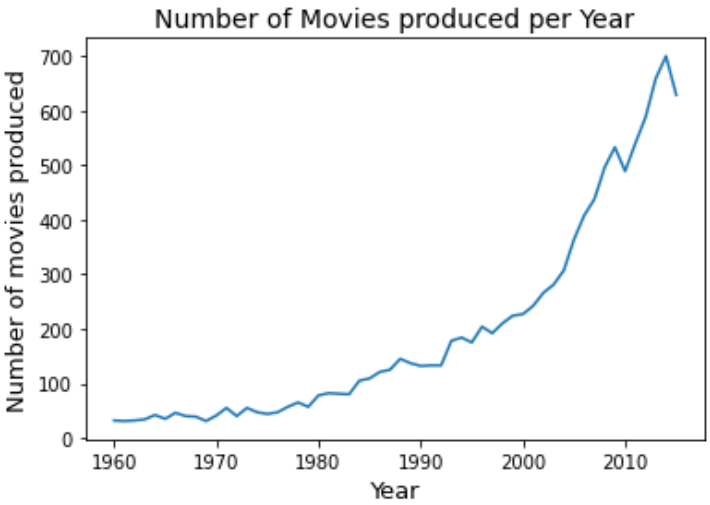

Figure 3.

4. What is the trend in the revenue over the years?

The earlier years had low revenue and the years between 2000 and 2005 had massive increase in revenue

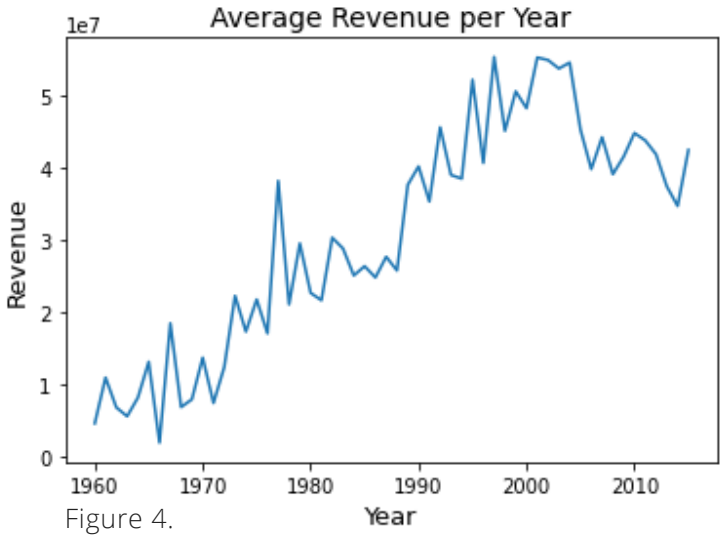

Figure 4.

5. What is the relationship between revenue and popularity?

There is a positive correlation between revenue and popularity. This means that popular movies get higher revenue

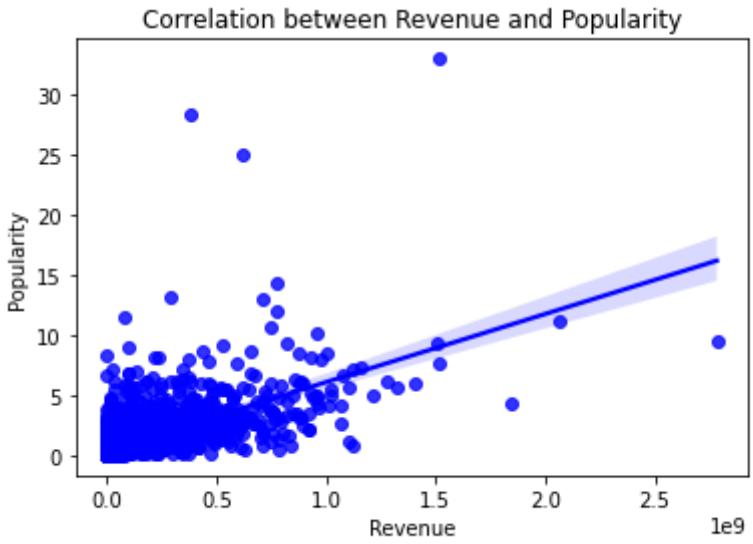

Figure 5.

6. What is the relationship between revenue and budget?

The earlier years had low revenue and the years between 2000 and 2005 had massive increase in revenue

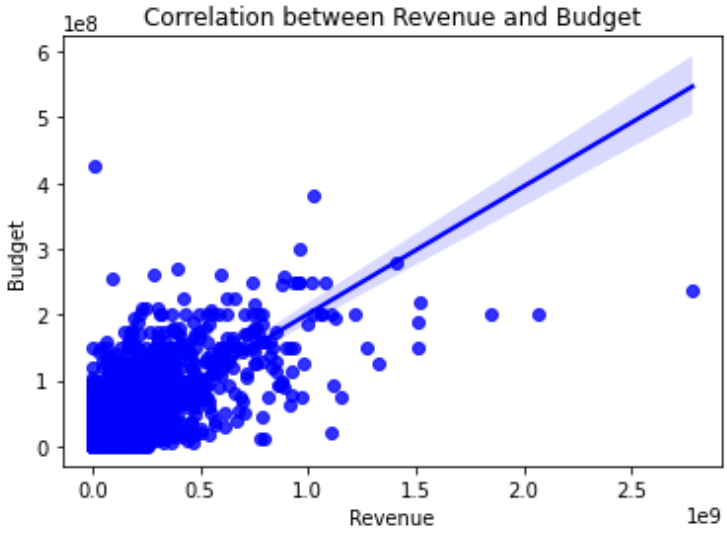

Figure 6.

7. Which genre has more releases?

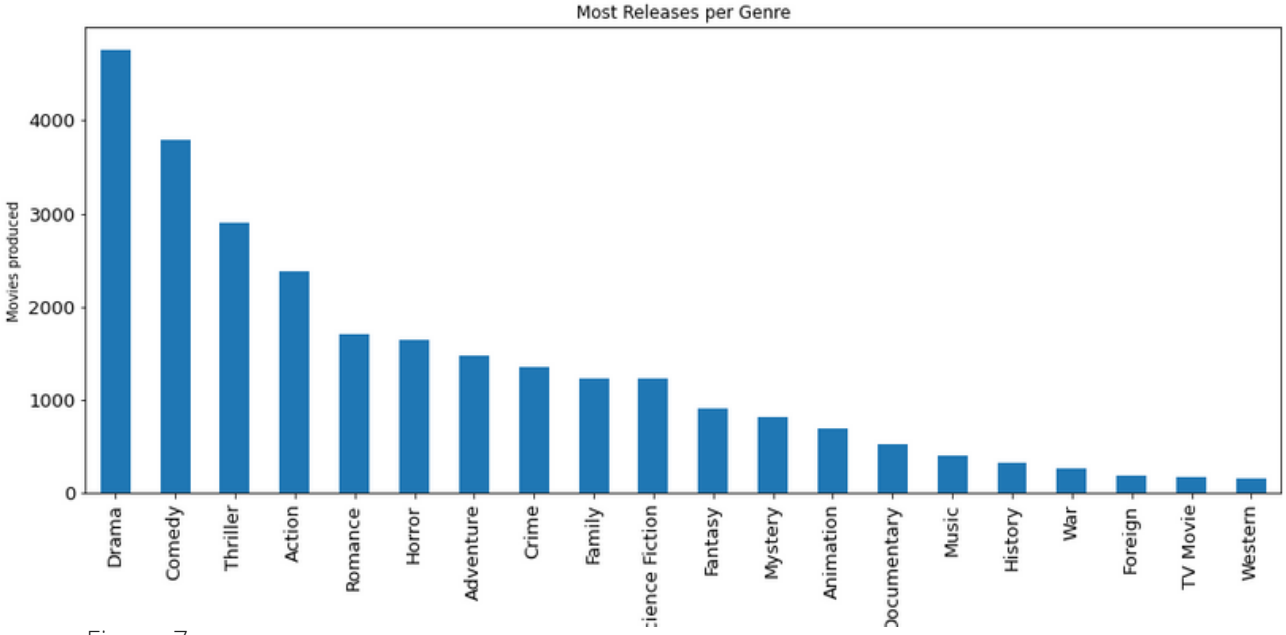

Figure 7.

More drama movies have been produced over ther years, folowed by comedy and thriller. This could also mean that drama movies are popular.

8. What is the relationship between revenue,popularity,runtime and budget?

There is a strong positive correlation between popularity, budget and revenue. This means that there is a high chance that movies with higher budgets earn more revenue and are more popular.

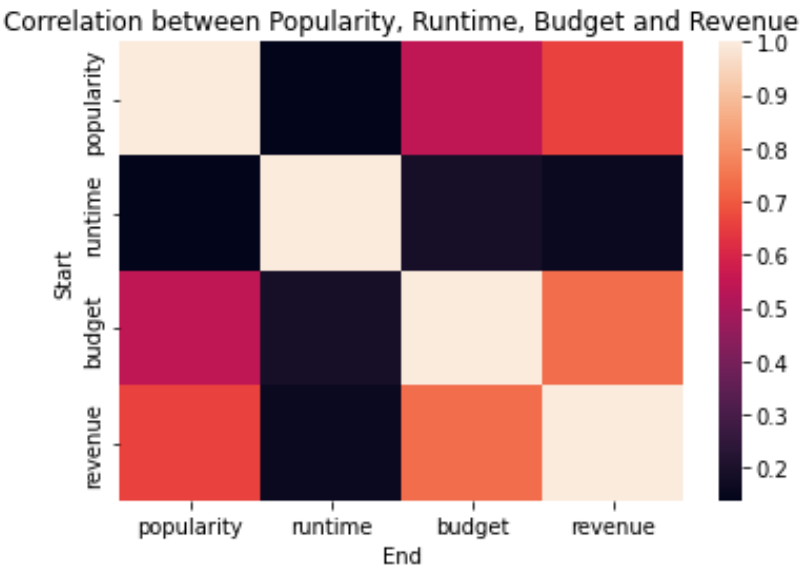

Figure 8.

CONCLUSION

1. Drama is the most popular genre also having the most releases.
2. The runtime of movies has been reducing over the years.
3. Budget is positively related to revenue.
4. Drama, comedy, thriller, action, and romance movies are the top five genres with the most releases.

LIMITATION

Missing values in the dataset hindered my analysis and I had to drop some columns with null values. Some columns with zero values also created some false correlations and this made me drop some of the columns that could have been handy in the analysis.

The budget and the revenue columns had zero values until the 50% percentile and I replaced them with the mean. The data frame has a better distribution but this could only be a representative of the zero values and could be close or far away from the correct values if they could be found.

REFERENCES

https://numpy.org/doc/stable/reference/generated/numpy.arange.html
https://matplotlib.org/3.5.0/gallery/images_contours_and_fields/image_annotated_heatmap.html
https://matplotlib.org/3.5.0/tutorials/introductory/pyplot.html
https://seaborn.pydata.org/generated/seaborn.heatmap.html
Udacity course content