Movies (TMDB) Analyzation

A. Introduction

This report consists of what we found in the movie dataset, the link for the dataset is shown below:
https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies/data

The data consists of 24 columns, some of them are:

1. id
2. title
3. vote_average
4. vote_count
5. status
6. release_date
7. revenue
8. runtime
9. adult (is it for adult or not)
10. backdrop_path
11. budget
12. etc.

The columns that will be used are:

Id, title, release_date, year, genres, budget, revenue, popularity, production companies, production countries.

B. Goals

The goal of this analysis is to find trends from the dataset from the year 2000, some of which are

1. Total Revenue made by movies
2. Best movies by the revenues
3. Best ROI
4. Revenue made by movies every year
5. Etc.

C. Data cleaning

As shown in chapter A, there are a lot of columns in this dataset, so we need to remove/drop them because we have no need for it.

df = df.drop(columns=["overview", "tagline", "poster_path", "keywords", "homepage", "backdrop_path","imdb_id"])

after dropping the unused columns next, we need to filter the movies, since we want to find trends, we remove the unreleased movies from the dataset. After we remove the movies, we dropped the status column.

```
#checking status

df = df.drop(df[df["status"]!="Released"].index)

df=df.drop(columns=["status"])
```

next step of this data cleaning step would be to remove rows that have 0 or no values in the main columns needed and remove duplicates.

```
#cleaning NaN Valued Rows

df = df.dropna(subset=["title", "release_date","genres"])

df = df.dropna(subset=["production_companies", "production_countries", "spoken_languages"])

#Checking duplicates

#cleaning dupes

df = df.drop_duplicates()
```

last step would be changing data type. Only one column has the wrong data type, the column in date/release_date. And then I take the year of the release_date as a release_year/year

```
#transform wrong type data

df['release_date'] = pd.to_datetime(df['release_date'])

df.info()

df['year'] = df['release_date'].dt.year
```

D. KPI

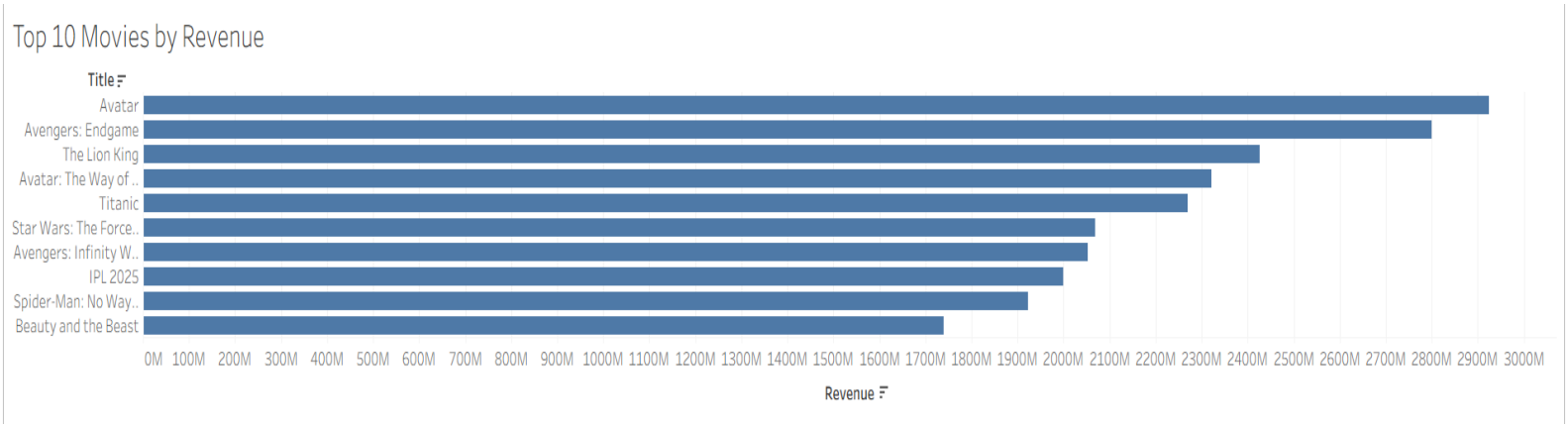
1. Top 5 Selling Movies
   This KPI is to find the best movies in terms of revenue
   ```
   top_selling_movie = df.sort_values(by='revenue', ascending=False).head(5)
   print(f"top selling movies : \n{top_selling_movie[['title','revenue']]}")
   ```

   The top 5 bestselling movies are

ID	Title	Revenue
3	Avatar	$ 2.923.706.026
15	Avengers: Endgame	$ 2.800.000.000

282	Avatar: The Way of Water	$ 2.320.250.281
56	Star Wars: The Force Awakens	$ 2.068.223.624
6	Avengers: Infinity War	$ 2.052.415.039

The top 10 are shown in the chart below



The chart is self-explanatory; the best movies are those that many have watched and many knows of the title, basically overall a popular show.

2. Best movies in terms of ROI

   This part explains and shows which movies have the best ROI (Return of Investment). ROI is calculated by:
   (Revenue – Budget)/Budget

```
df = df[df['budget']>0]
df = df[df['revenue']>0]
df['ROI'] = (df['revenue'] - df['budget'])/df['budget']
top_roi_movie = df.sort_values(by='ROI', ascending=False).head(5)
print(f"top Return of Investment movies are : \n{top_roi_movie[['title','ROI']]}")
```

   to find the best ROI we need to filter it, the movies that will be calculated are those that have budget>0 & Revenue >0. Here are the top 5 best movies in terms of ROI

id	title	ROI	Budget
784593	La vie d'un kiwi	2.000000e+07	5
597743	The NeedyMonster	5.333332e+06	150
258495	24: Redemption	3.398367e+06	4
59258	Between Us	2.755583e+06	1
426919	the evil FNAF attack	1.999999e+06	1

   object

As seen on the list on top, most of the movies with the best ROIs are indie movies, this happens because of how much budget is used in production, and when it was released, although it seems the top 5 has weird budget used, because it is in one to three digits. All we can say is the top 5 are outliers in terms of budget used.

3. Bestselling genre for movies

This section tells the revenue for each genre, here are the top 5 overall bestselling genre from the year 2000,

Adventure 241.136.866.610

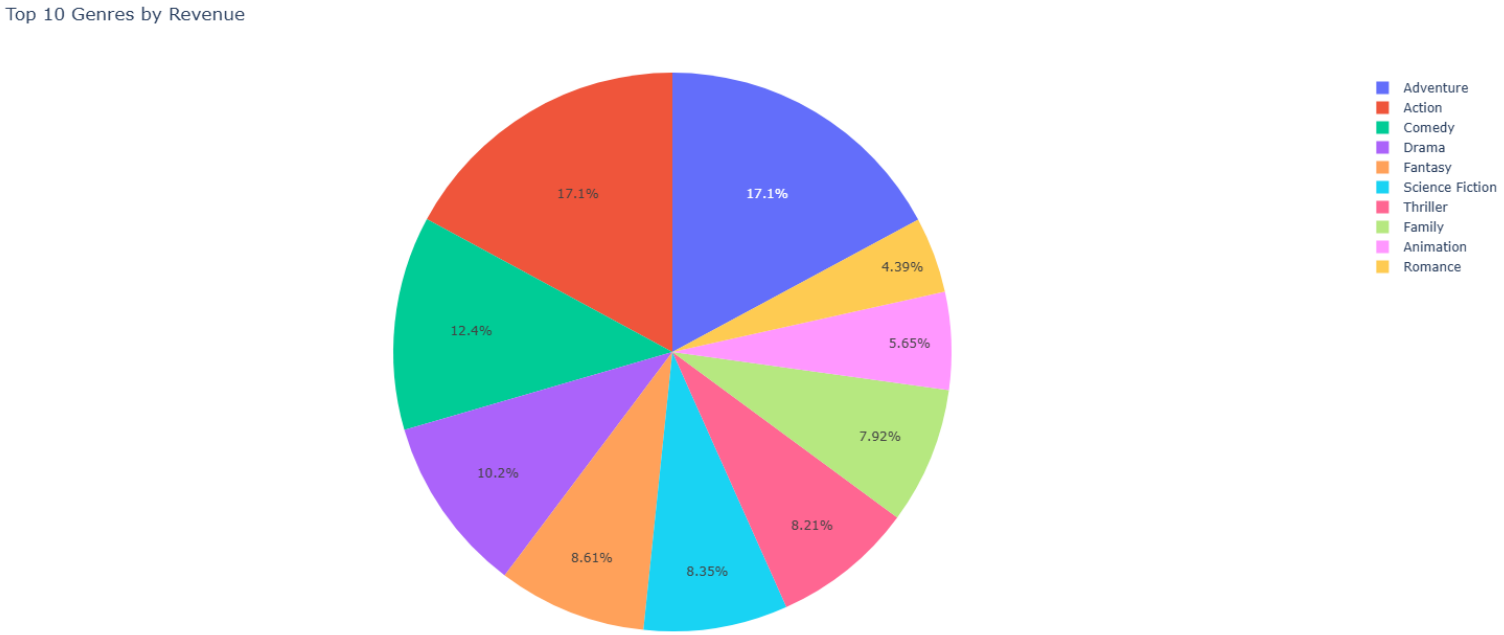
Action 240.787.562.850

Comedy 174.235.952.393

Drama 143.993.678.948

Fantasy 121.091.015.557

How much percentage does every genre hold in terms of revenue would be shown in the chart below



As shown in the chart on top, adventure and action are leading with the percentage of 17.1%, although adventure holds the highest in terms of value. Followed by comedy with 12.4% and then drama with 10.2%. This percentage

shows that people, or at least movies watcher tend to enjoy and love to watch action or adventure movies. When producers want to release a movie, and there are a lot of movies released at the same time, people would want to watch the movies with these two genres the most, that is adventure or action. This shown in the best revenue made by movies, the top 5 are movies with the genre adventure or action.

4. Production companies ranking in revenue
   This section will explain which production companies or studios have the highest revenue overall and some of the reasons for it.
   The highest revenue held by a company overall from 2000 to 2025 is currently Warner Bros. Pictures with the amount of revenue made $ 59.941.714.049
   As shown in section 3 of KPI the most genre generated the highest revenue are Action or Adventure, this company made a lot of movies title with this genre the list is shown below.

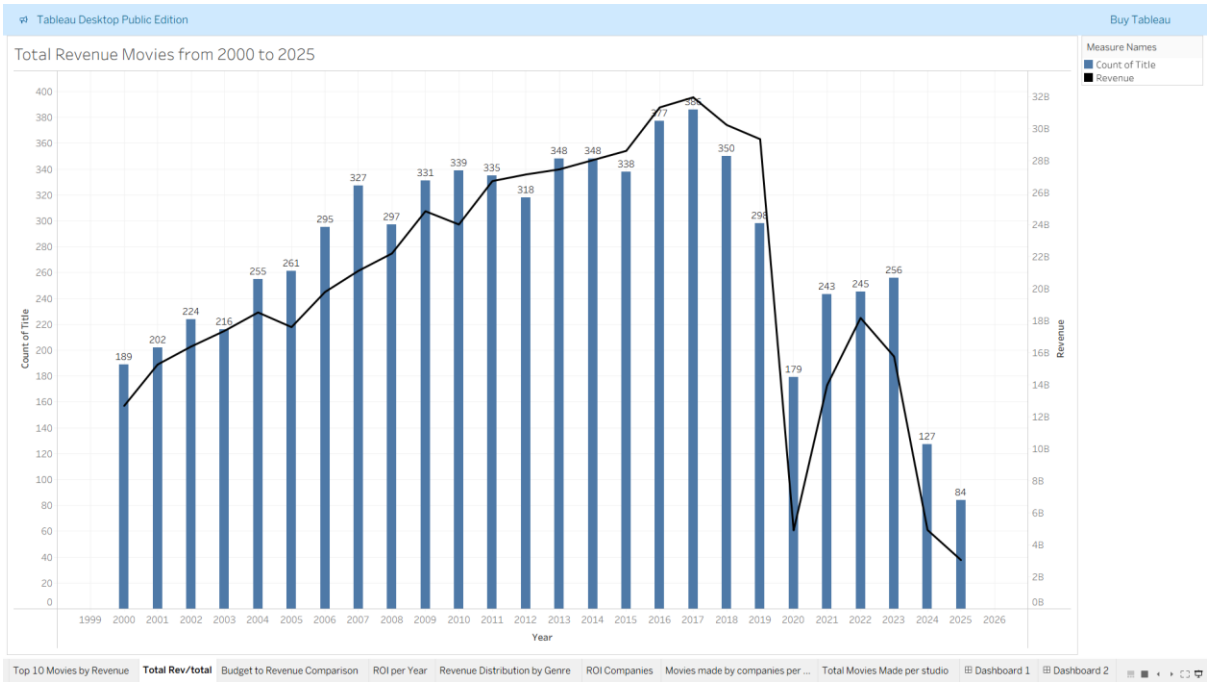
Movies made by companies per genre

Producti..	Genres (Movies Gen..	Count of Genres (Movies Genre Cleaned.Csv)	Count of Title
Warner Bros. Pictures	Action	206,0	206,0
	Adventure	149,0	149,0
	Animation	28,0	28,0
	Comedy	201,0	201,0
	Crime	152,0	152,0
	Documentary	4,0	4,0
	Drama	283,0	283,0
	Family	51,0	51,0
	Fantasy	71,0	71,0
	History	34,0	34,0
	Horror	41,0	41,0
	Music	25,0	25,0
	Mystery	65,0	65,0
	Romance	104,0	104,0
	Science Fiction	75,0	75,0
	Thriller	185,0	185,0
	War	27,0	27,0
	Western	20,0	20,0

A bit of disclaimer some titles may hold 2 of the genres at the same time, because there are not any movies that have only one genre.

E. Overall Analysis

This chapter will explain overall analysis that have not been explained in the last section.



The Overall Revenue made by movies in the span of 25 year are distributed as shown in the chart above. The black line shows the revenue made every year, while the blue bar shows the total movies released at that year. The highest revenue generated by movies was done in 2017 with the amount of movie title released also at the highest.

The chart above shows the comparison between revenue generated per year and budget used to produce films per year; the red line shows the budget while the black line shows the revenue. As the chart above shows, from the year 2000 until 2013, most movie industry was not making enough revenue for profit, it increased in 2014 until 2018. This happens because some movies are a lot more popular than the

others, for example Avengers: Endgame and Avengers: Infinity War that was released in the year 2019 and 2018. From the year 2019 until 2025 the movie industry had a decline in terms of profit again, but the budget was also declining at the same time. This happened because in the year 2019 COVID hit, and most cinemas were closed; people shifted from watching movies in the cinemas to streaming online. A lot of platforms skyrocketed in terms of popularity this year. Another reason for this is because a lot of companies withheld their productivity due to work from home policy, that policy is the reason why in the year 2019 until 2022 there was a decline in movie released.