TMDB Movie Data Analysis

Harsh Kumar Gupta

Mentored by : Sir Ashish Rajput

CONTENTS

INTRO & OBJECTIVE

Helping a movie company to analyse the type of movies and genres that perform well in cinema.

BUSINESS OBJECTIVE

1. Display the movie categories that have a budget greater than $220,000.
2. Display the movie categories where the revenue is greater than $961,000,000.
3. Remove the rows with value 0 from both the budget and revenue columns.
4. List the top 10 movies with the highest revenues and the top 10 movies with the least budget.
5. How are popularities of movies related with the movie budgets?
6. Identify and display the names of all production companies along with the number of times they appear in the dataset.
7. Display the names of the top 25 production companies based on the number of movies they have produced in descending order of the number of movies produced.
8. Sort the data in descending order based on revenue and filter the top 500 movies and perform Outlier analysis
9. Identify and display the names of the movies along with their run times for those movies that have above average runtime

Data Preprocessing

```
data=pd.read_csv('DS1_C8_V3_ND_Sprint2_Data Analysis Using Python_Dataset.csv')
data.head(2)
```

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity	production_compan
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577	[{"name": "Ingenic Film Partners", " 28
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	139.082615	[{"name": "Walt Disr Pictures", "id": 2}, {

```
json_columns=['genres','keywords','production_companies','production_countries','spoken_languages']

for col in json_columns:
    data[col]=data[col].apply(lambda x: str(x)[1:-1])
```

```
def extract_names(col_val):
    name_list = json.loads(f'[{col_val}]')
    names = [col_val['name'] for col_val in name_list]
    return ', '.join(names)

for col in json_columns:
    data[f'{col}_name'] = data[col].apply(extract_names)
    data[f'{col}_name']=data[f'{col}_name'].apply(lambda x:x.split(', '))
```

```
data.drop(columns=json_columns,inplace=True)
```

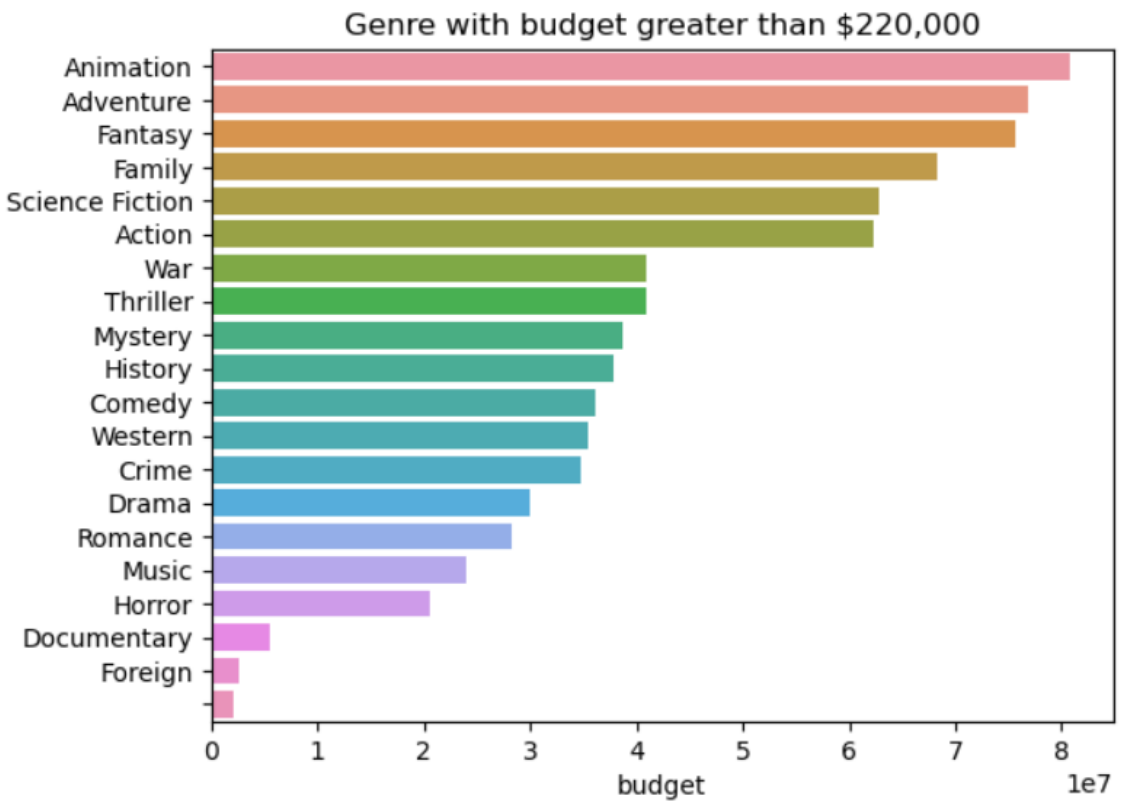
```
data[['title','genres_name']]
```

	title	genres_name
0	Avatar	[Action, Adventure, Fantasy, Science Fiction]
1	Pirates of the Caribbean: At World's End	[Adventure, Fantasy, Action]
2	Spectre	[Action, Adventure, Crime]
3	The Dark Knight Rises	[Action, Crime, Drama, Thriller]
4	John Carter	[Action, Adventure, Science Fiction]
...	...	...
4773	Clerks	[Comedy]
4788	Pink Flamingos	[Horror, Comedy, Crime]
4792	Cure	[Crime, Horror, Mystery, Thriller]
4796	Primer	[Science Fiction, Drama, Thriller]
4798	El Mariachi	[Action, Crime, Thriller]

3 Display the movie categories that have a budget greater than $220,000.

```
data_genre_flaten=data.explode('genres_name')
t3_nocondition=pd.pivot_table(index='genres_name',values='budget',data=data_genre_flaten,aggfunc='mean')
t3=t3_nocondition[t3_nocondition.budget>220000].sort_values('budget',ascending=False)
```

```
1 plt.title('Genre with budget greater than $220,000')
2 sns.barplot(x=t3.budget,y=t3.index)
3 plt.xlabel('budget')
4 plt.ylabel('')
5 plt.show()
```



BEFORE

	original_title	genres_name
0	Avatar	[Action, Adventure, Fantasy, Science Fiction]

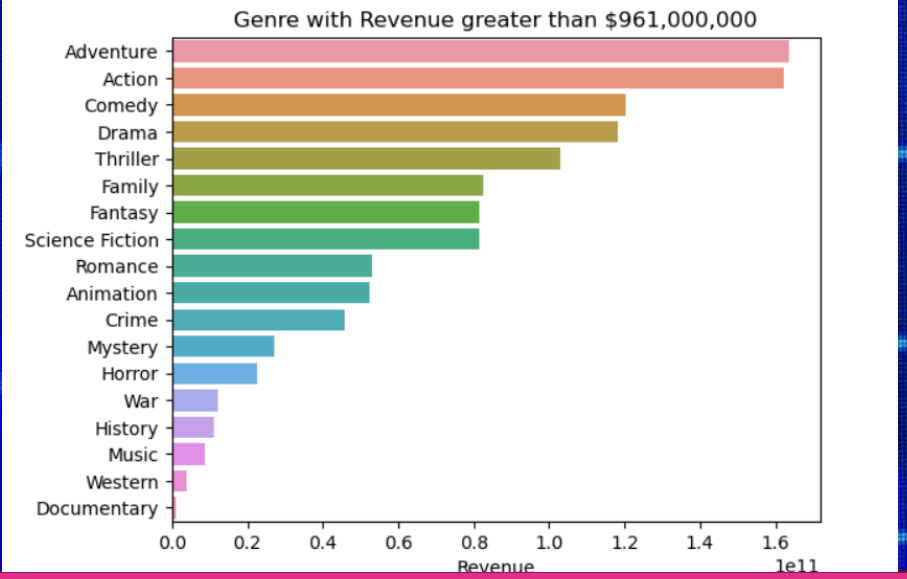
AFTER FLATEN

original_title	genres_name
Avatar	Action
Avatar	Adventure
Avatar	Fantasy
Avatar	Science Fiction

1. There are total of 20 genre that have budget greater than $220,000
2. Animation genre has the highest budget

4 Display the movie categories where the revenue is greater than $961,000,000.

```
plt.title('Genre with Revenue greater than $961,000,000')
sns.barplot(x=t4.revenue,y=t4.index)
plt.xlabel('Revenue')
plt.ylabel('')
plt.show()
```



1. Adventure Genre has generated the highest revenue
2. Documentary genre has generated the least revenue

5 Remove the rows with value 0 from both the budget and revenue columns.

```
data = data[(data['budget'] != 0) ]
data = data[(data['revenue'] != 0) ]
```

1. The revenue 0 and budget 0 movies were dropped

6 List the top 10 movies with the highest revenues and the top 10 movies with the least budget.

Highest revenue

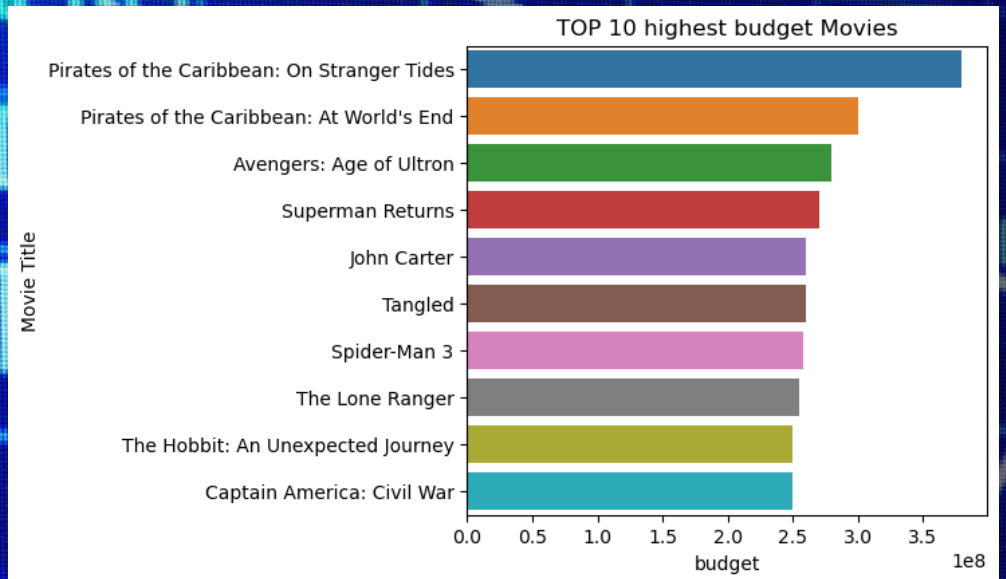
```
top_10_movies=data[['original_title','budget']]\
.sort_values('budget',ascending=False).nlargest(10,columns='budget')
top_10_movies
```

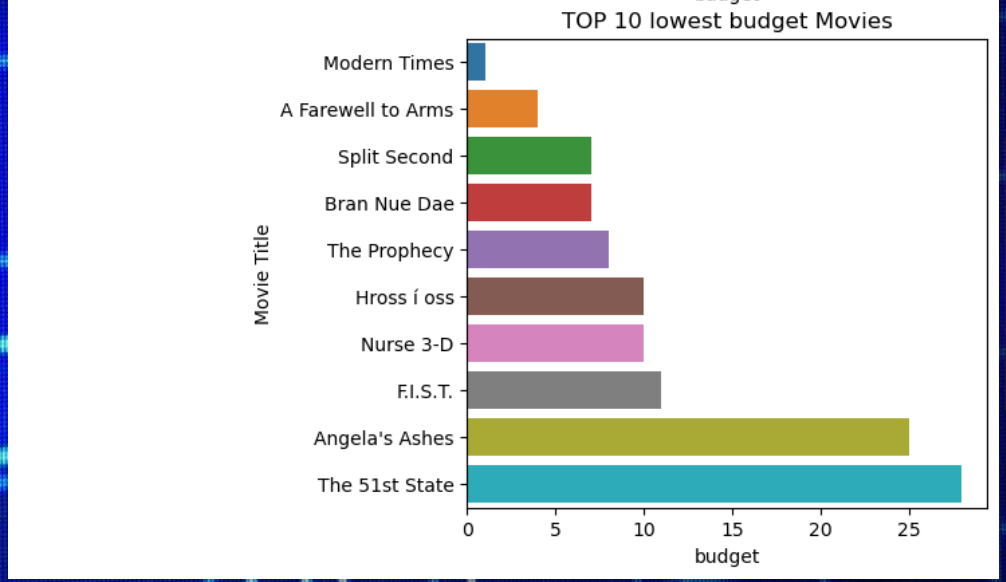
	original_title	budget
17	Pirates of the Caribbean: On Stranger Tides	380000000
1	Pirates of the Caribbean: At World's End	300000000
7	Avengers: Age of Ultron	280000000
10	Superman Returns	270000000
4	John Carter	260000000
6	Tangled	260000000
5	Spider-Man 3	258000000
13	The Lone Ranger	255000000
98	The Hobbit: An Unexpected Journey	250000000
26	Captain America: Civil War	250000000

Least revenue

```
top_10lowest_movies=data[['original_title','budget']]\
.sort_values('budget',ascending=False).nsmallest(10,columns='budget')
top_10lowest_movies
```

	original_title	budget
4238	Modern Times	1
3611	A Farewell to Arms	4
3372	Split Second	7
3419	Bran Nue Dae	7
4608	The Prophecy	8
3131	Hross í oss	10
3137	Nurse 3-D	10
2933	F.I.S.T.	11
1912	Angela's Ashes	25
1771	The 51st State	28

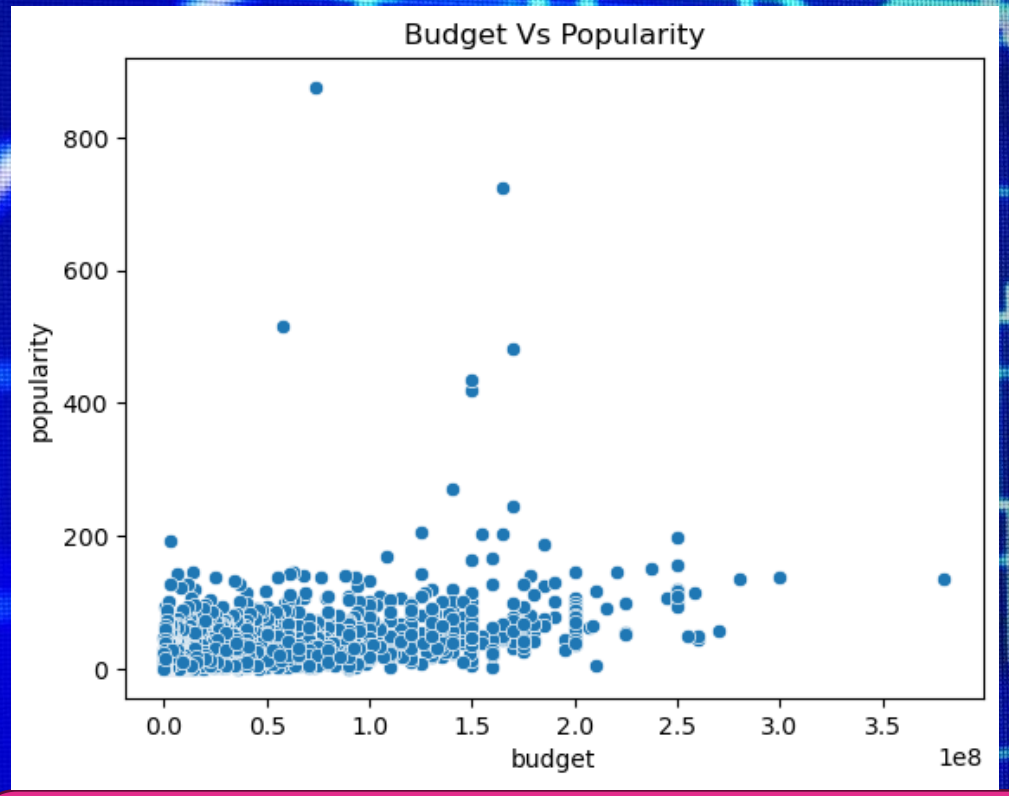




Pirates of the Caribbean is the movie with the highest budget

Modern Times is the movie with the lowest budget

7 How are popularities of movies related with the movie budgets?



1. There is no correlation between the budget and popularity
2. Most of the movies are made in the budget between $ 0 to $ 20000000

8 Display the names of all production companies along with the number of times they appear in the dataset.

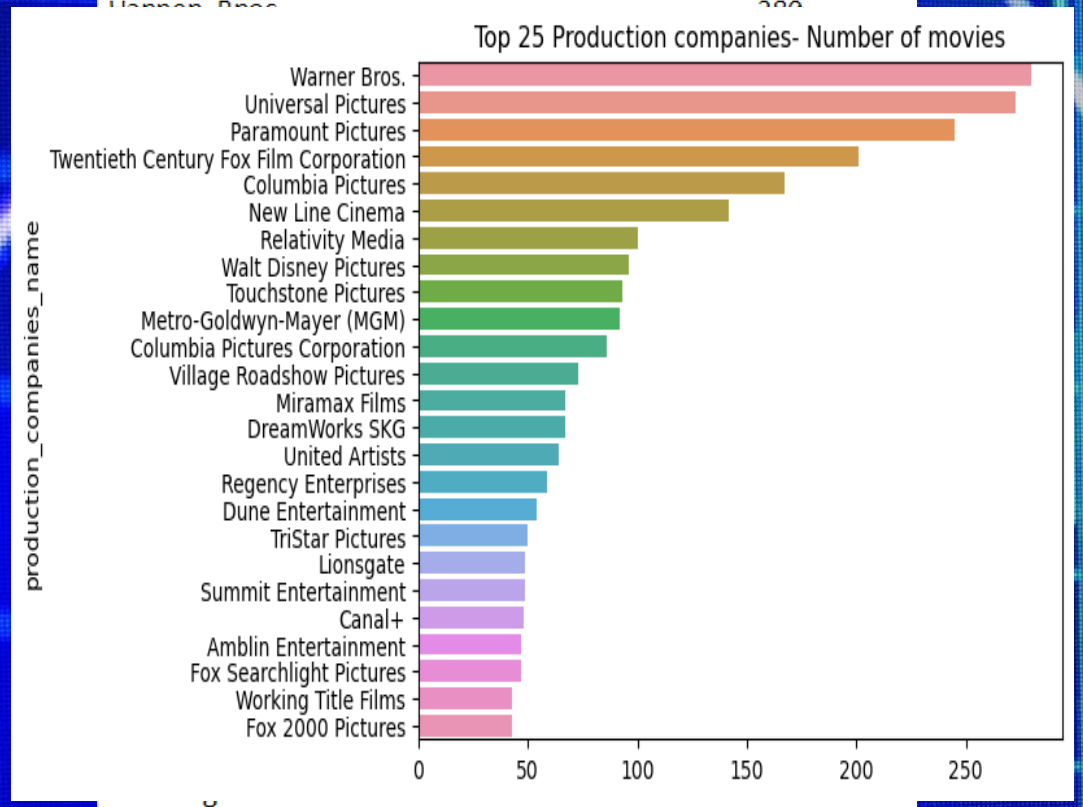
```
1 data_productionname_flatten=data.explode('production_companies_name')
2 t8=data_productionname_flatten.production_companies_name.value_counts()
3 t8
```

```
production_companies_name
Warner Bros.                                280
Universal Pictures                          273
Paramount Pictures                          245
Twentieth Century Fox Film Corporation      201
Columbia Pictures                           167
                                           ...
Micro Fusion 2003-2                           1
HW Two                                        1
Unfinished Films                              1
Infinity Features Entertainment               1
Daiei Studios                                 1
Name: count, Length: 3571, dtype: int64
```

1. There are total 3571 movie production companies
2. Warner Bros. has produced 280 movies which is the highest in the dataset

9 Display the names of the top 25 production companies based on the number of movies

```
production_companies_name
Warner Bros.                                    280
```




```
Fox 2000 Pictures                               43
Name: count, dtype: int64
```

1. Warner Bros has produced most number of movies

10 Sort the data in descending order based on revenue and filter the top 500 movies. (Perform outlier analysis)

```
def remove_outliers(df, columns):
    df_no_outliers = df.copy()

    for col in columns:

        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        LowerBound = Q1 - 1.5 * IQR
        UpperBound = Q3 + 1.5 * IQR

        df_no_outliers = df_no_outliers[(df_no_outliers[col] > LowerBound) & (df_no_outliers[col] < UpperBound)]

    return df_no_outliers
```

```
t10_no_outliers=remove_outliers(t10,columns)
```

```
t10_no_outliers.shape
```

```
(466, 20)
```

All the outliers were removed and at the end we were left with 466 movies

11 Display the movie categories where the revenue is greater than $961,000,000.

```
1 t10_no_outliers[t10_no_outliers.runtime>t10_no_outliers.runtime.mean()][['title','runtime']]
```

	title	runtime
113	Harry Potter and the Order of the Phoenix	138.0
8	Harry Potter and the Half-Blood Prince	153.0
330	The Lord of the Rings: The Two Towers	179.0
233	Star Wars: Episode I - The Phantom Menace	136.0
675	Jurassic Park	127.0
...	...	...
521	The Terminal	128.0
397	It's Complicated	121.0
1744	Knocked Up	129.0
717	Jack Reacher	130.0
714	Collateral	120.0

220 rows × 2 columns

There are220 movies whose runtime is greater than the average average runtime

CONCLUSION

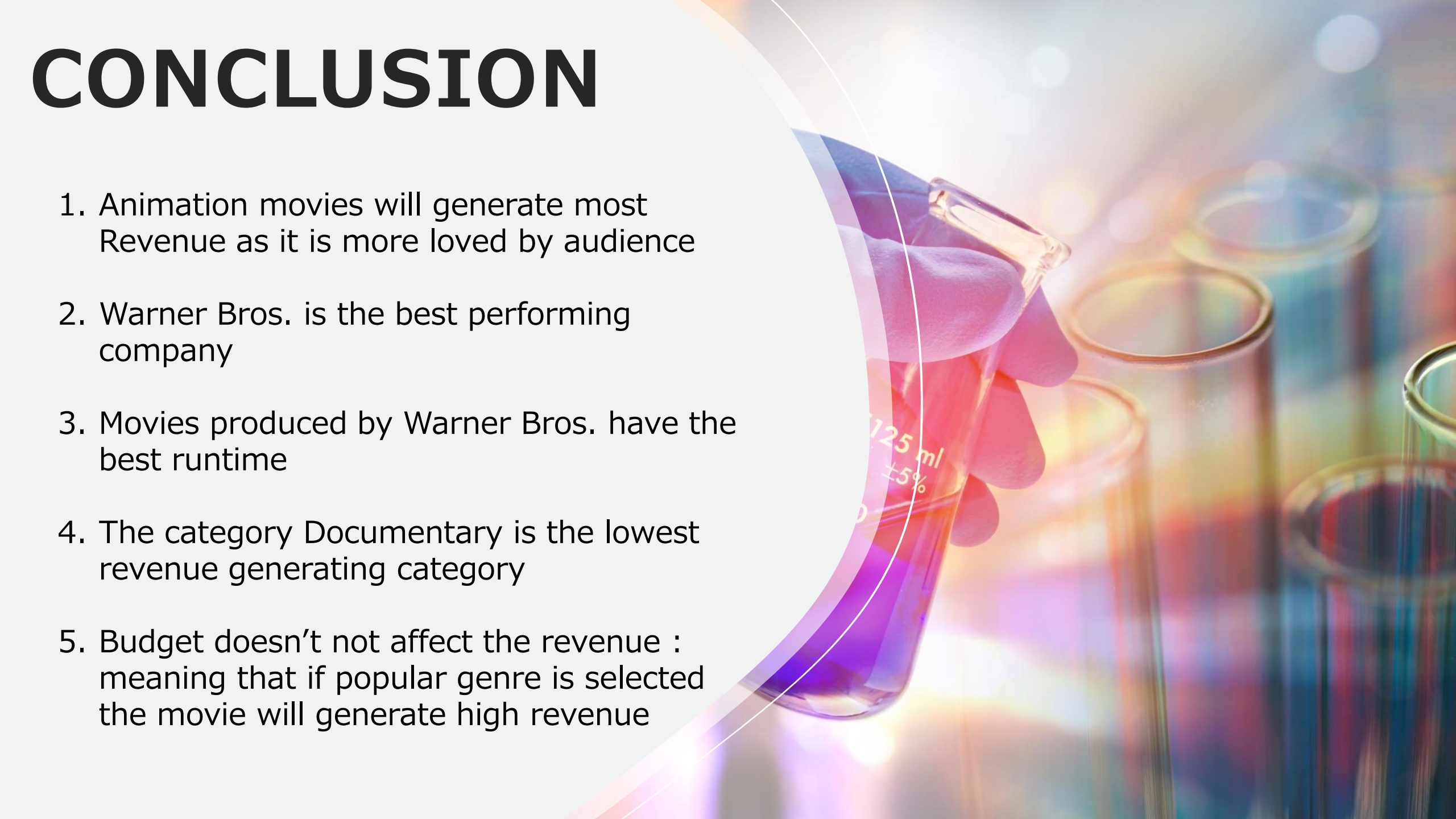
1. Animation movies will generate most Revenue as it is more loved by audience

2. Warner Bros. is the best performing company

3. Movies produced by Warner Bros. have the best runtime

4. The category Documentary is the lowest revenue generating category

5. Budget doesn’t not affect the revenue : meaning that if popular genre is selected the movie will generate high revenue



THANK YOU