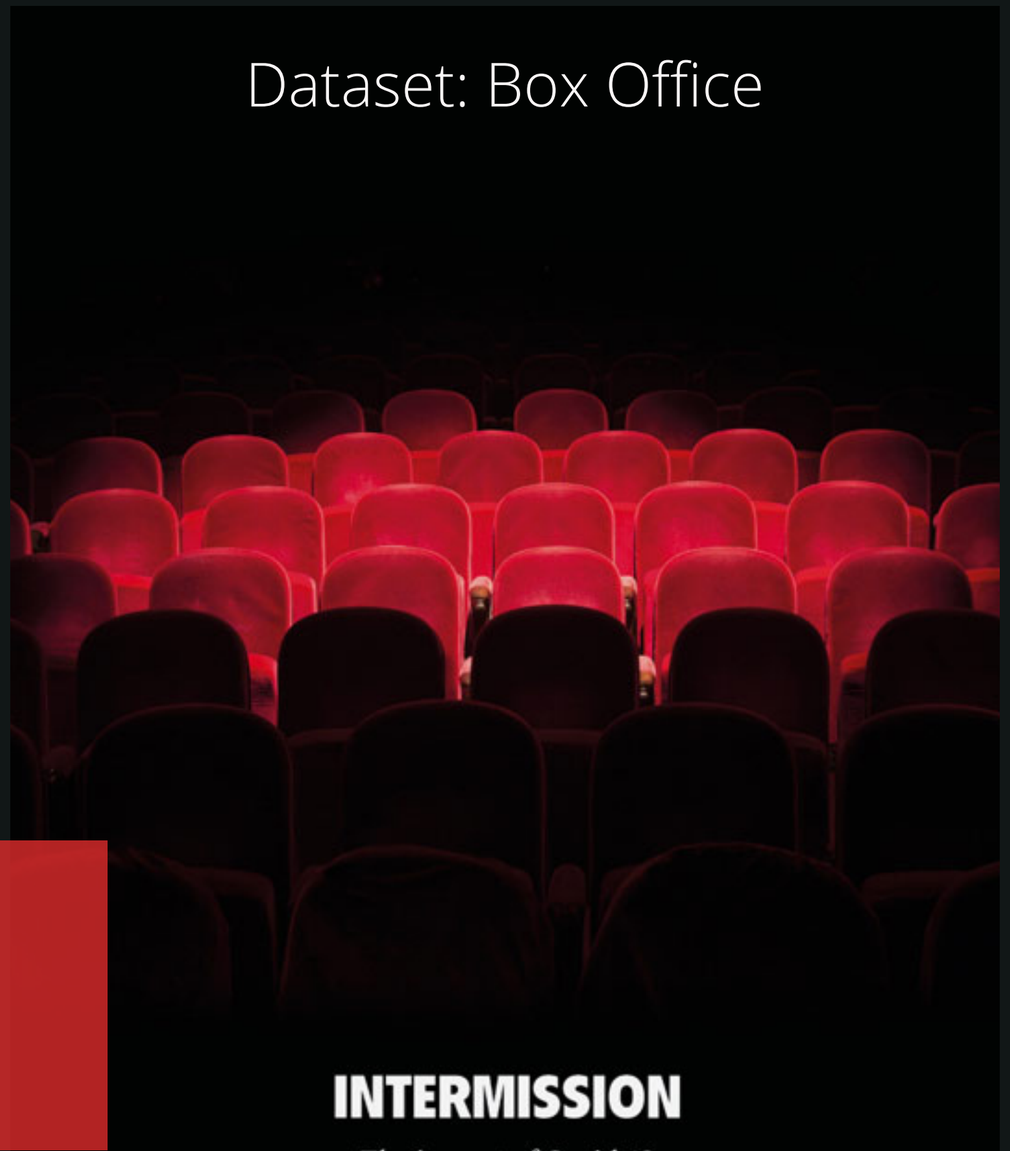
Group 8 Project Report

Movie Revenue Prediction

Dataset: Box Office



By:
Ashutosh Bhatt
Bunty Kher
Joseph Dankwah
Kenneth Akinbode
SushmithPatel

Introduction

- **Define the business problem. Why do you think what you work on is important?**
  - **What do you plan to achieve with your project?**
  - **What is the primary goal of the project, and how will it help the business?**
  - **What are the implications of the project?**
- **Describing Dataset**
  - **Where has the data been gathered?**
  - **What are the input and output variables?**
- **Descriptive Analysis**
  - **Number of columns and rows**
  - **Percent of missing values by columns**
  - **Means, medians, modes, correlations, variance, outliers**
  - **Data visualizations such as scatter plots or correlation plots**
- **Data Preprocessing**
  - **Variable treatment**
  - **Dropping columns and rows when necessary**
  - **Imputing missing values**
  - **Transforming variables if necessary**
  - **Variable selection**
- **Fitting models**
  - **Fitting multiple models with varying parameters**
  - **Summarize performance indicators of models**
  - **Selecting the best model using a test data set**

Business Problem

Making movies is as big a business venture as investing millions into a stock or into a project that is expected to bring revenue over a period. For production houses and studios, a clear picture of their Return on Investment helps them seek financing and investments, allocate resources and so on.

- Many studios are not aware of how to properly estimate or project their revenue which in turn causes many losses when their initial revenue projections fall short.

Why do you think what you work on is important?

- The problem we seek to address is the Knowledge gap between investment and profit generation that exists within the movie production space.

What is the primary goal of the project, and how will it help the business?

The primary goal of the project is to enable the business to project revenue and profits of a movie to be produced based on specific variables including the estimated budget as observed in data we have gathered from over 2400 movies over the past 20 years.

For our projections, we wish to research the following:

- Top 10 actors bring in the most profit
- Top 10 movie genres bring in the most profit
- Top 10 movie director brings in the most profit
- Predict how much profit a studio will make by employing any set of actors and directors for the top genres
- Correlation between budget and Profit (ROI)

WHAT ARE THE IMPLICATIONS OF THE PROJECT?

- On completion of this project, the business will save a lot of money that ordinarily may have been expended in production if we realise that the success of the movie (in terms of finances) is not dependent necessarily on a big budget.

- Additionally, this project will also help the business in their casting process since we may be able to tie box office revenues to specific actors or actors of a certain calibre

- The film industry can save a lot of money by using this tools as this can help them to hire a trending actor and director as well as which genre is trending among the audience.

Dataset Description

Where has the data been gathered?

The dataset comprises data collected from over 2400 movies over a 20 year period. They were scrapped from the box-office mojo website and have data about movies that were released within the US over the above period. This dataset was downloaded from Kaggle for the purpose of this project

NO OF ROWS: 2476
NO OF COLUMNS: 25

	movie_id	movie_title	movie_year	director	writer	producer	composer	cinematographer
0	tt0118589	Glitter	2001	Vondie Curtis-Hall	Cheryl L. West	Laurence Mark	Terence Blanchard	Geoffrey Simpson
1	tt0120630	Chicken Run	2000	Peter Lord	Peter Lord	Peter Lord	Harry Gregson-Williams	Simon Jacobs
2	tt0120667	Fantastic Four	2005	Tim Story	Mark Frost	Avi Arad	John Ottman	Oliver Wood
3	tt0120679	Frida	2002	Julie Taymor	Hayden Herrera	Lindsay Flickinger	Elliot Goldenthal	Rodrigo Prieto
4	tt0120681	From Hell	2001	Albert Hughes	Alan Moore	Jane Hamsher	Trevor Jones	Peter Deming
...	...	...	...	...	...	...	...	...
2471	tt8688634	21 Bridges	2019	Brian Kirk	Adam Mervis	Chadwick Boseman	Alex Belcher	Paul Cameron
2472	tt8946378	Knives Out	2019	Rian Johnson	Rian Johnson	Ram Bergman	Nathan Johnson	Steve Yedlin
2473	tt9024106	Unplanned	2019	Chuck Konzelman	Abby Johnson	Chris Jones	Stephen Blake Kanicka	Drew Maw
2474	tt9134216	Playing with Fire	2019	Andy Fickman	Dan Ewen	Todd Garner	Nathan Wang	Dean Semler
2475	tt9173418	Brahms: The Boy II	2020	William Brent Bell	Stacey Menear	Matt Berenson	Brett Detar	Karl Walter Lindenlaub

2476 rows × 25 columns

source:- https://www.kaggle.com/kerneler/starter-boxofficemojo-dataset-18c3edbb-3/data

What are the input and output variables?

INPUT VARIABLES

- Domestic
- Worldwide
- International
- Mpaa
- Budget
- Profit or loss

OUTPUT VARIABLES

Return of Investment - we predict ROI of movie

Descriptive Analysis

Descriptive Statistics of Numerical Variables

```
mdata.describe()
```

	movie_year	budget	domestic	international	worldwide	roi
count	2476.000000	2.476000e+03	2.476000e+03	2.476000e+03	2.476000e+03	2.476000e+03
mean	2008.872375	5.055923e+07	6.680113e+07	8.809556e+07	1.548967e+08	1.043375e+08
std	5.634442	4.955858e+07	8.273985e+07	1.508960e+08	2.255849e+08	1.921207e+08
min	2000.000000	2.200000e+02	8.000000e+03	9.990000e+02	7.962400e+04	-1.107663e+08
25%	2004.000000	1.600000e+07	1.696048e+07	7.741622e+06	2.967356e+07	3.507790e+06
50%	2009.000000	3.500000e+07	3.979211e+07	3.385308e+07	7.607761e+07	4.136070e+07
75%	2014.000000	6.900000e+07	8.305228e+07	9.787582e+07	1.805715e+08	1.217376e+08
max	2020.000000	3.560000e+08	9.366622e+08	2.029931e+09	2.797801e+09	2.553439e+09

Descriptive Statistics of Categorical Variables

```
mdata.describe(include='all')
```

	movie_title	movie_year	director	writer	producer	composer	cinematographer	cast	budget	domestic	international	worldv
count	2476	2476.000000	2476	2476	2476	2476	2476	2476	2.476000e+03	2.476000e+03	2.476000e+03	2.476000e
unique	2465	NaN	1209	1645	1182	591	673	2460	NaN	NaN	NaN	N
top	The Girl with the Dragon Tattoo	NaN	Clint Eastwood	Luc Besson	Jason Blum	Christophe Beck	Dean Semler	Bradley Cooper,Zach Galifianakis,Ed Helms,Just...	NaN	NaN	NaN	N
freq	2	NaN	16	15	39	132	115	2	NaN	NaN	NaN	N
mean	NaN	2008.872375	NaN	NaN	NaN	NaN	NaN	NaN	5.055923e+07	6.680113e+07	8.809556e+07	1.548967e
std	NaN	5.634442	NaN	NaN	NaN	NaN	NaN	NaN	4.955858e+07	8.273985e+07	1.508960e+08	2.255849e
min	NaN	2000.000000	NaN	NaN	NaN	NaN	NaN	NaN	2.200000e+02	8.000000e+03	9.990000e+02	7.962400e
25%	NaN	2004.000000	NaN	NaN	NaN	NaN	NaN	NaN	1.600000e+07	1.696048e+07	7.741622e+06	2.967356e
50%	NaN	2009.000000	NaN	NaN	NaN	NaN	NaN	NaN	3.500000e+07	3.979211e+07	3.385308e+07	7.607761e
75%	NaN	2014.000000	NaN	NaN	NaN	NaN	NaN	NaN	6.900000e+07	8.305228e+07	9.787582e+07	1.805715e

PERCENT OF MISSING VALUES BY COLUMNS

```
In [97]: # Percent of missing values by columns
         percent_missing = mdata.isnull().sum() * 100 / len(mdata)
         missing_value_mdata = pd.DataFrame({'column_name': mdata.columns,
                                             'percent_missing': percent_missing})
```

```
In [98]: print(missing_value_mdata)
```

	column_name	percent_missing
movie_id	movie_id	0.000000
movie_title	movie_title	0.000000
movie_year	movie_year	0.000000
director	director	0.000000
writer	writer	0.161551
producer	producer	0.201939
composer	composer	2.948304
cinematographer	cinematographer	3.634895
main_actor_1	main_actor_1	0.000000
main_actor_2	main_actor_2	0.000000
main_actor_3	main_actor_3	0.000000
main_actor_4	main_actor_4	0.000000
cast	cast	0.000000
budget	budget	0.000000
domestic	domestic	0.000000
international	international	0.000000
worldwide	worldwide	0.000000
mpaa	mpaa	2.504039
run_time	run_time	0.000000
genre_1	genre_1	0.000000
genre_2	genre_2	7.310178
genre_3	genre_3	30.129241
genre_4	genre_4	63.529887
Genres	Genres	0.000000
link	link	0.000000

MEANS:

What Genre and combination of Genre manage to get how much return of investment mean?

```
In [111]: #What Genre and combination of Genre manage to get how much return of investment mean
          mdata.groupby('Genres')['roi'].mean()

Out[111]: Genres
          Action                                9.795106e+06
          Action,Adventure                      5.547914e+08
          Action,Adventure,Animation,Comedy     2.835372e+08
          Action,Adventure,Animation,Drama      1.842806e+07
          Action,Adventure,Animation,Family     1.980470e+08
                                                    ...
          Horror,Thriller                       5.848955e+07
          Music                                -1.002419e+07
          Mystery,Thriller                      8.035742e+07
          Romance,Sci-Fi,Thriller,              7.766938e+07
          Sci-Fi,Thriller                       4.751001e+07
          Name: roi, Length: 476, dtype: float64
```

What is the average earnings of movies by year?

```
In [112]: # average earnings of movies by year
          mdata.groupby('movie_year')['roi'].mean().sort_values(ascending=False)

Out[112]: movie_year
          2017    1.870040e+08
          2019    1.855537e+08
          2015    1.676209e+08
          2018    1.550124e+08
          2012    1.531792e+08
          2014    1.469892e+08
          2016    1.415889e+08
          2013    1.303519e+08
          2007    1.065589e+08
          2008    9.965447e+07
```

MEDIAN AND MODE

Impute the missing numerical variables and categorical variables using the median & mode respectively

```
for i in numerical_var:
    mdata[i] = mdata[i].fillna(mdata[i].median(), inplace = False)

for i in categorical_var:
    mdata[i] = mdata[i].fillna(mdata[i].mode()[0], inplace = False)

mdata.head()
```

	movie_id	movie_title	movie_year	director	writer	producer	composer	cinematographer	main_actor_1	main_actor_2	...	international
0	tt0118589	Glitter	2001	Vondie Curtis-Hall	Cheryl L. West	Laurence Mark	Terence Blanchard	Geoffrey Simpson	Mariah Carey	Eric Benét	...	997259
1	tt0120630	Chicken Run	2000	Peter Lord	Peter Lord	Peter Lord	Harry Gregson-Williams	Simon Jacobs	Mel Gibson	Julia Sawalha	...	118000000
2	tt0120667	Fantastic Four	2005	Tim Story	Mark Frost	Avi Arad	John Ottman	Oliver Wood	Ioan Gruffudd	Michael Chiklis	...	178839854
3	tt0120679	Frida	2002	Julie Taymor	Hayden Herrera	Lindsay Flickinger	Elliot Goldenthal	Rodrigo Prieto	Salma Hayek	Alfred Molina	...	30413474
4	tt0120681	From Hell	2001	Albert Hughes	Alan Moore	Jane Hamsher	Trevor Jones	Peter Deming	Johnny Depp	Heather Graham	...	42955549

5 rows × 25 columns

```
print(missing_value_mdata)
```

	column_name	percent_missing
movie_title	movie_title	0.0
movie_year	movie_year	0.0
director	director	0.0
writer	writer	0.0
producer	producer	0.0
composer	composer	0.0
cinematographer	cinematographer	0.0
cast	cast	0.0
budget	budget	0.0
domestic	domestic	0.0
international	international	0.0
worldwide	worldwide	0.0
mpaa	mpaa	0.0
run_time	run_time	0.0
Genres	Genres	0.0

CORRELATIONS:

We used a correlation between budget and ROI.

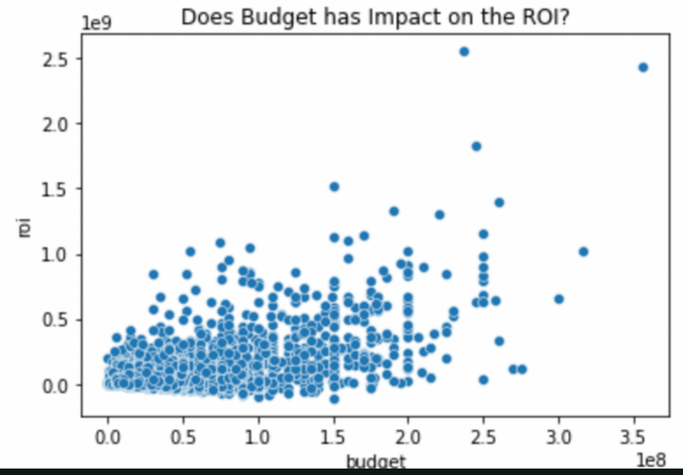
```
In [101]: #Using correlation between budget and roi

          corr_budget_roi = mdata["budget"].corr(mdata['roi'])
          print(corr_budget_roi)

          0.6050763390104473

In [102]: sns.scatterplot(x="budget", y="roi", data=mdata)
          plt.title("Does Budget has Impact on the ROI?")

Out[102]: Text(0.5, 1.0, 'Does Budget has Impact on the ROI?')
```



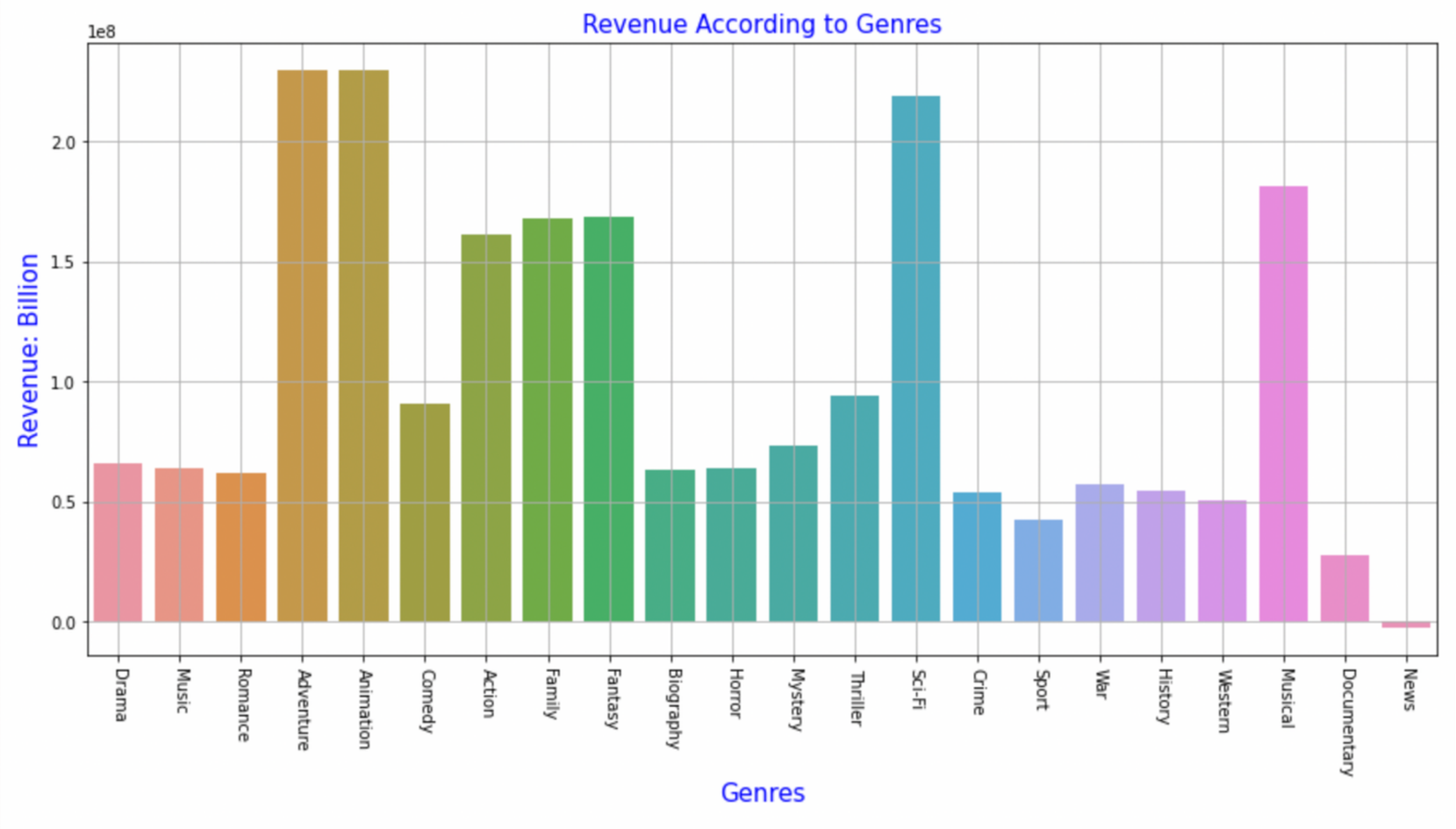

```
[mdata.corr( )['roi'].sort_values(ascending=False) ]

[roi                1.000000
 worldwide          0.984585
 international      0.964570
 domestic           0.925282
 budget             0.605076
 movie_year         0.193229
 Name: roi, dtype: float64]
```

WITH THE HELP OF A SCATTERPLOT,
WE CAN DIRECTLY SEE THAT THE BUDGET AND THE ROI ARE CORRELATED WITH 60% AND WORLDWIDE COLLECTION IS 98%

Data visualizations

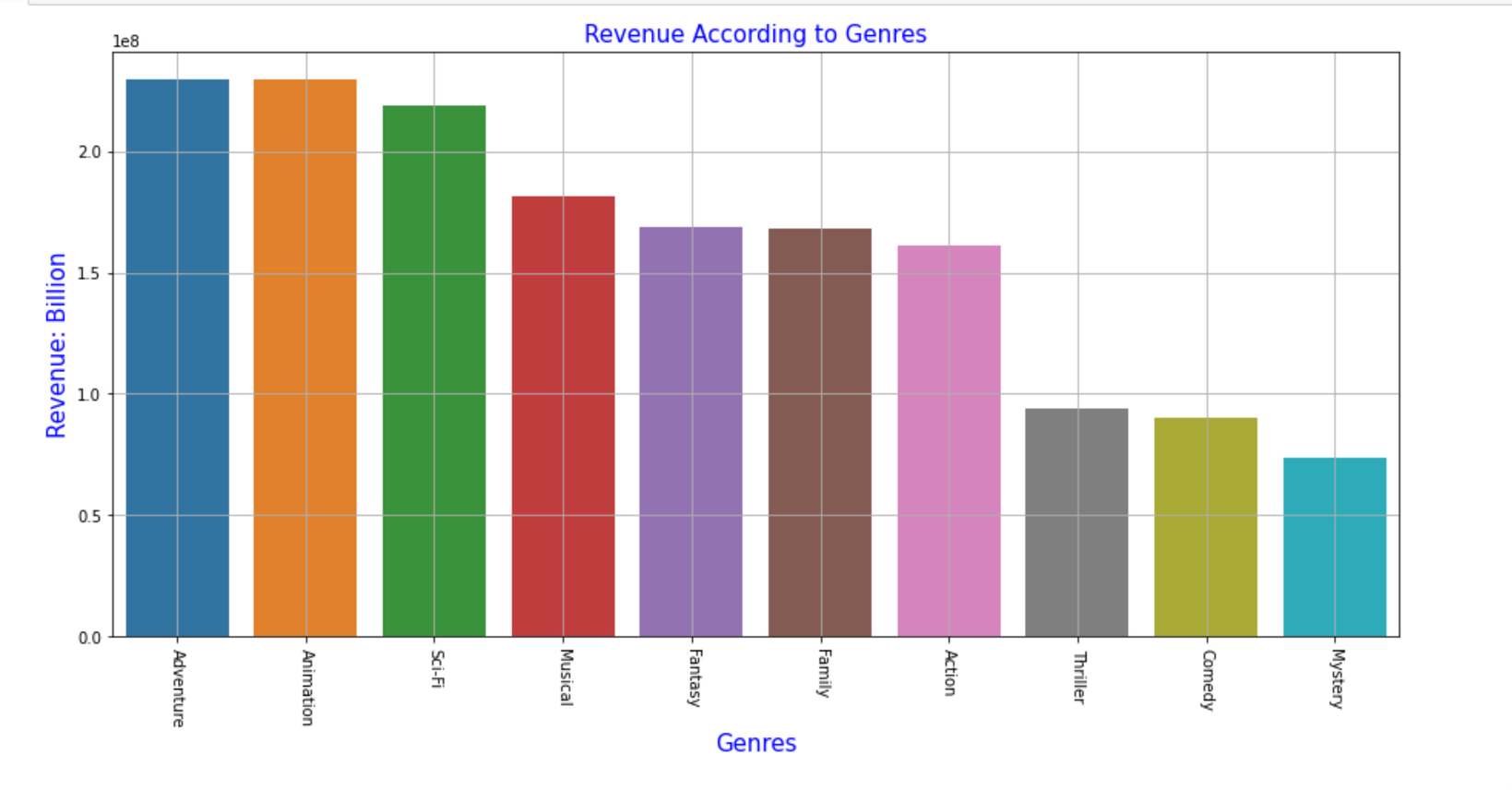
Here, it shows average revenue from every genre of the movies from past 20 years.



WE CAN CLEARLY SEE THAT ACCORDING TO OUR DATASET, ADVENTURE, ANIMATION AND SCIFI ARE THE TOP MONEY MAKING GENRE.

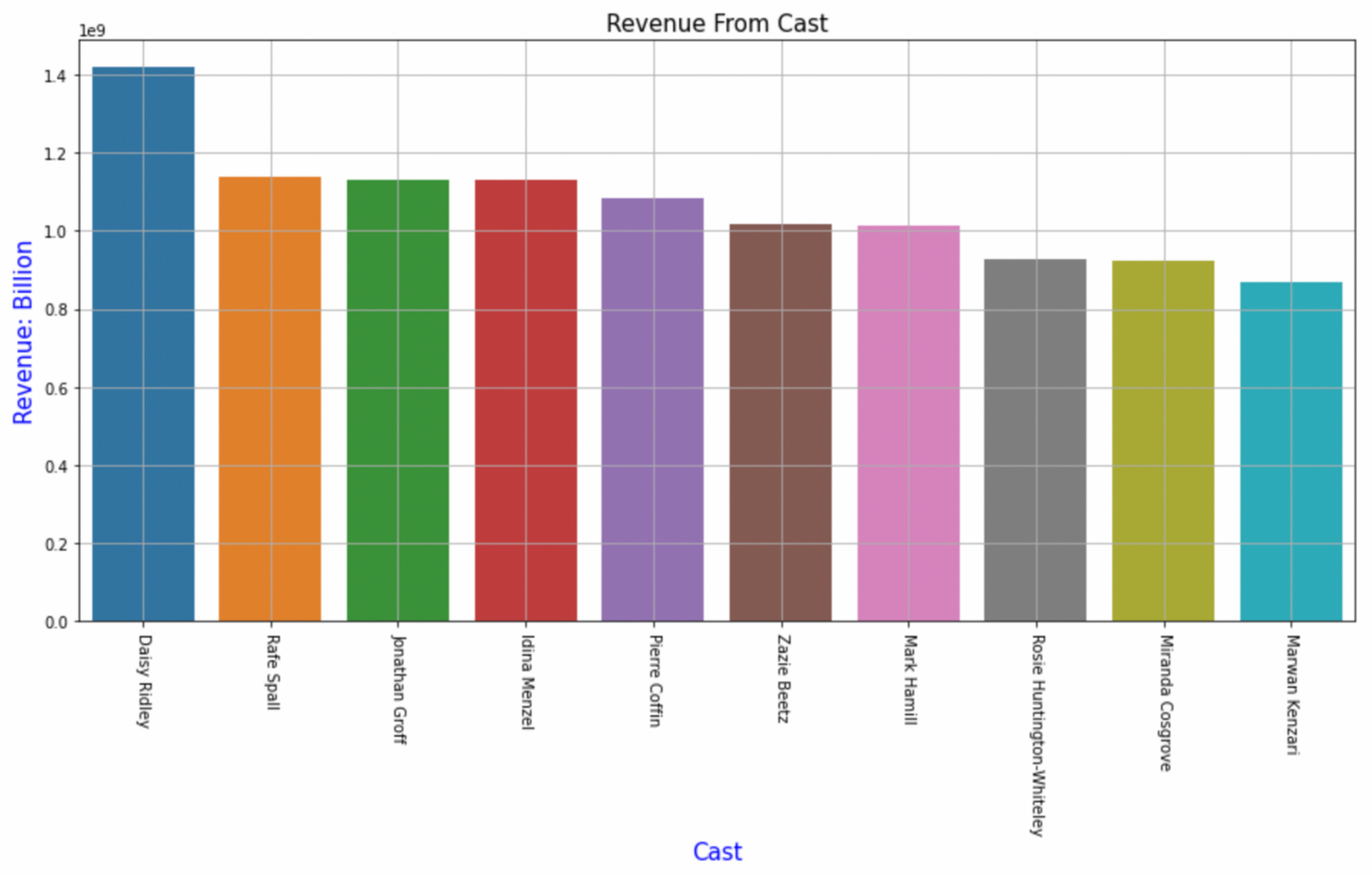
Data visualizations

The Bar plot shows Top 10 Genres who brought on an average high revenue



Data visualizations

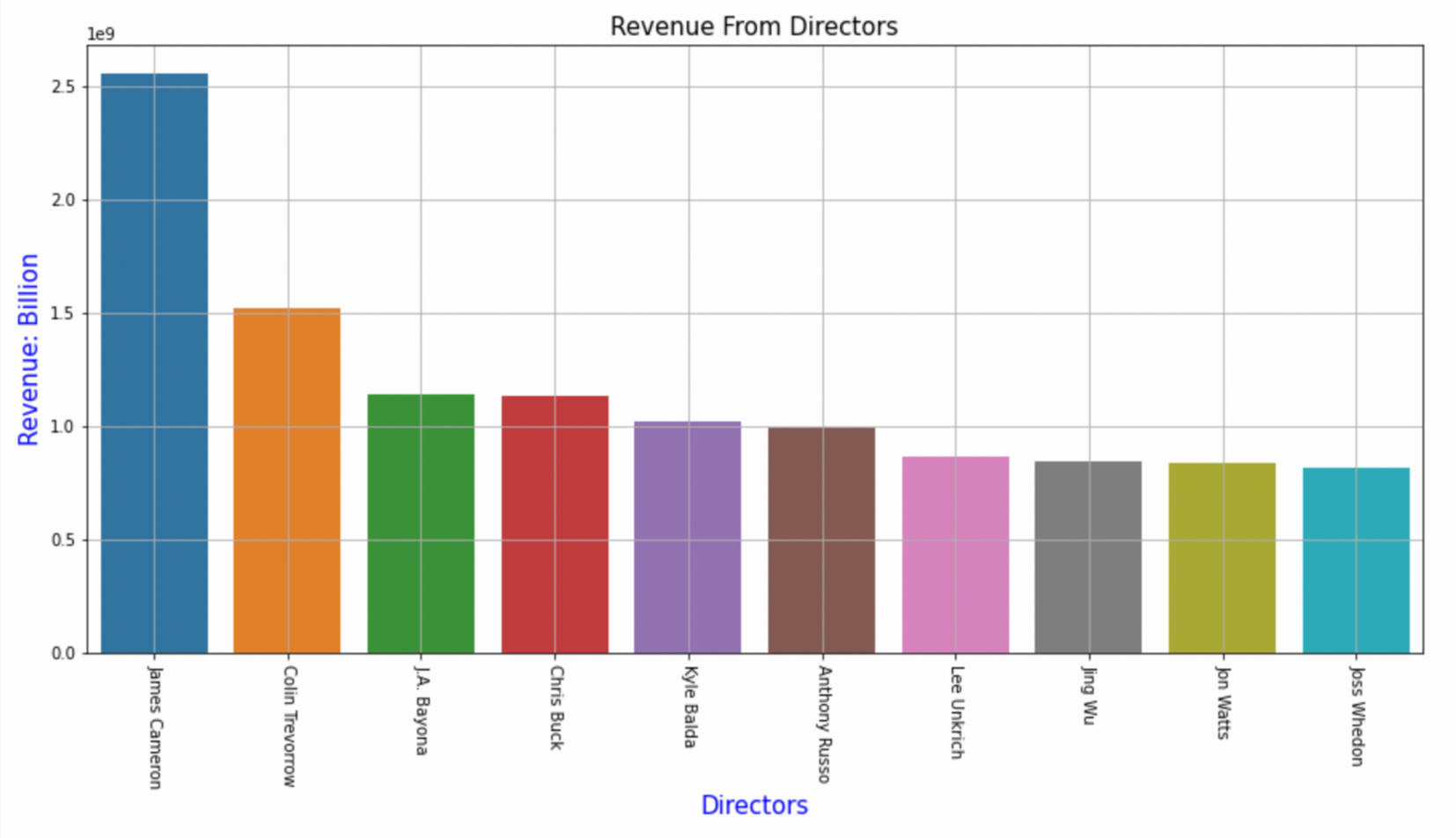
Here, it shows average revenue from Top 10 Actors & Actresses of the movies from past 20 years.



WE CAN CLEARLY SEE THAT ACCORDING TO OUR DATASET, DAISY RIDLEY'S MOVIE ARE MAKING MORE REVENUE THAN OTHER CASTS.

Data visualizations

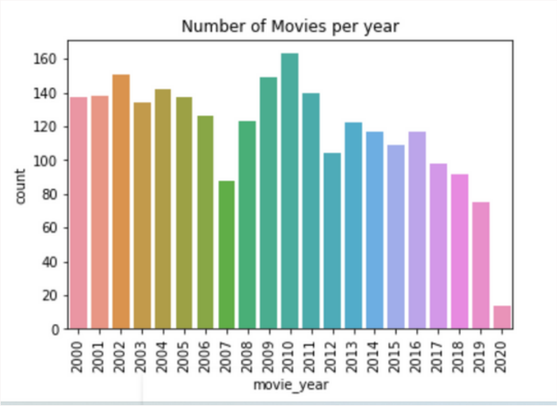
Here, it shows average revenue from Top 10 Directors of the movies from past 20 years.

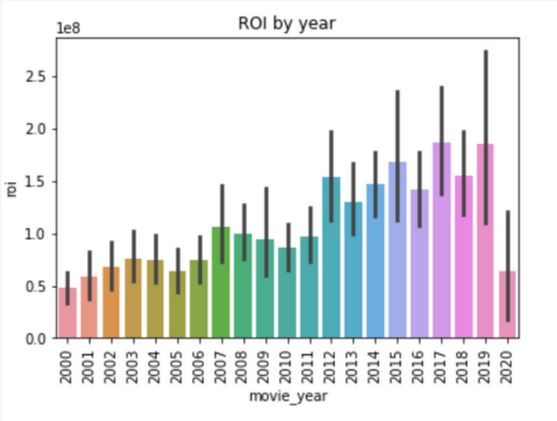


WE CAN CLEARLY SEE THAT ACCORDING TO OUR DATASET, JAMES CAMERON'S MOVIE ARE MAKING MORE REVENUE THAN OTHER CASTS.

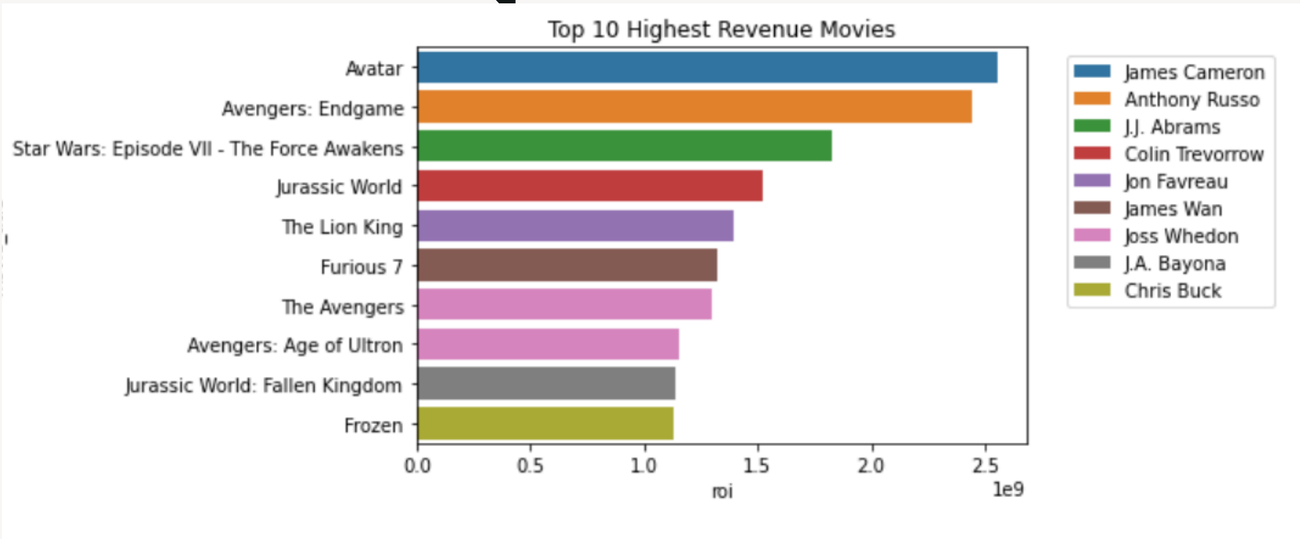
Data visualizations

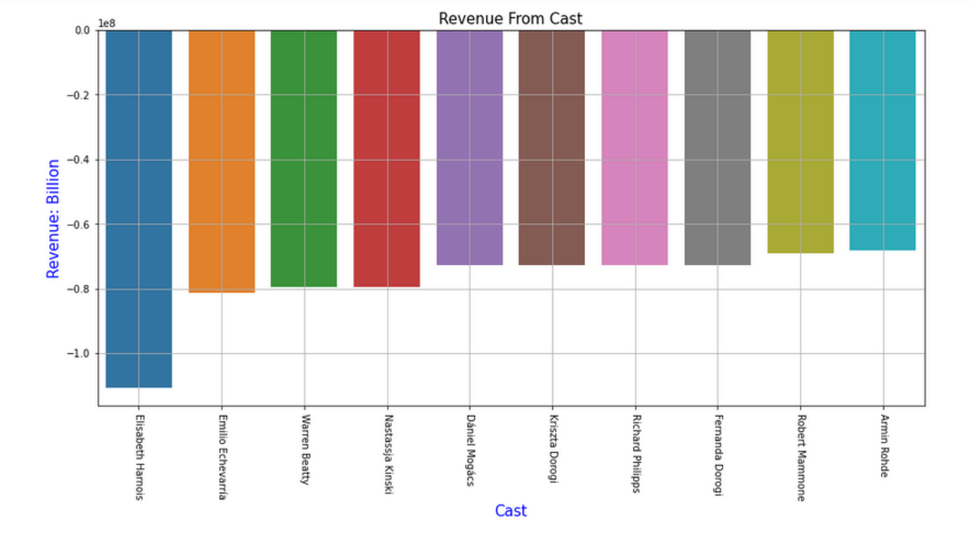
From the charts besides, we can clearly visualize the Number of movies per year and Return of Investment by year





From the below charts, we can clearly visualize the Top 10 Highest Revenue Movies and their Directors, Top 10 Actors/Actress whose average revenue from past 20 years is in the loss. So, Studios can determine that they should not go with those actors.





VARIANCE AND OUTLIERS:

Variance of the ROI Column.

```
In [114]: #variance of the roi column
          mdata.loc[:,'roi'].var()

Out[114]: 3.691035793480142e+16
```

VARIANCE IS THE EXPECTATION OF THE SQUARED DEVIATION OF A RANDOM VARIABLE FROM ITS POPULATION MEAN OR SAMPLE MEAN.

Data Preprocessing

Out[77]:

	movie_year	budget	domestic	international	worldwide	roi	loss	movie_title_10,000 BC	movie_title_102 Dalmatians	movie_title_10th & Wolf	...	Genres_Horror,Mystery
0	2001	22000000	4274407	997259	5271666	-16728334	1	0	0	0	...	
1	2000	45000000	106834564	118000000	224834564	179834564	0	0	0	0	...	
2	2005	100000000	154696080	178839854	333535934	233535934	0	0	0	0	...	
3	2002	12000000	25885000	30413474	56298474	44298474	0	0	0	0	...	
4	2001	35000000	31602566	42955549	74558115	39558115	0	0	0	0	...	

DROPPING COLUMNS AND ROWS WHEN NECESSARY:

We drop all the links, extra genre columns, actor columns and movie IDs.

```
In [87]: mdata.drop(['link'], axis=1, inplace=True)

In [88]: mdata.drop(['genre_1'], axis=1, inplace=True)

In [89]: mdata.drop(['genre_2'], axis=1, inplace=True)

In [90]: mdata.drop(['genre_3'], axis=1, inplace=True)

In [91]: mdata.drop(['genre_4'], axis=1, inplace=True)

In [92]: mdata.drop(['main_actor_1'], axis =1, inplace = True)

In [93]: mdata.drop(['main_actor_2'], axis =1, inplace = True)

In [94]: mdata.drop(['main_actor_3'], axis =1, inplace = True)

In [95]: mdata.drop(['main_actor_4'], axis =1, inplace = True)

In [96]: mdata.drop(['movie_id'], axis =1, inplace = True)

In [97]: mdata.head()
```

"WE ALREADY IMPUTED MISSING NUMERICAL VALUES BY USING MEDIAN AND MODE"

TRANSFORMING VARIABLES IF NECESSARY

We are using dummy variables for Loss and Profit from ROI (Return of Investment) Variable

```
In [125]: j=[]
          for i in mdata ["roi"]:
                  j.append("profit" if i>0 else"loss")
          mdata["pro-loss"] = j
```

```
In [126]: dummy = pd.get_dummies(mdata['pro-loss'])
          mdata = pd.concat((mdata, dummy), axis=1)
```

```
In [127]: mdata.head()
```

Out[127]:

matographer	cast	budget	domestic	international	worldwide	mpaa	run_time	Genres	roi	pro-loss	loss	profit
ffrey Simpson	Mariah Carey,Eric Benét,Max Beesley,Da Brat	22000000	4274407	997259	5271666	PG-13	1 hr 44 min	Drama,Music,Romance,	-16728334	loss	1	0
Simon Jacobs	Mel Gibson,Julia Sawalha,Phil Daniels,Lynn Fer...	45000000	106834564	118000000	224834564	PG-13	1 hr 24 min	Adventure,Animation,Comedy,Drama,	179834564	profit	0	1
Oliver Wood	Ioan Gruffudd,Michael Chiklis,Chris Evans,Jess...	100000000	154696080	178839854	333535934	PG-13	1 hr 46 min	Action,Adventure,Family,Fantasy	233535934	profit	0	1
Rodrigo Prieto	Salma Hayek,Alfred Molina,Geoffrey Rush,Mía Ma...	12000000	25885000	30413474	56298474	R	2 hr 3 min	Biography,Drama,Romance,	44298474	profit	0	1
Peter Deming	Johnny Depp,Heather Graham,Ian Holm,Robbie Col...	35000000	31602566	42955549	74558115	R	2 hr 2 min	Horror,Mystery,Thriller,	39558115	profit	0	1

Fitting multiple models with varying parameters

We are going to compare the 3 models

- Linear Regression
- Random Forest Regressor
- Random Forest Regressor with Twerked parameter (n_estimator = 10)

Random Forest Regressor

```
x=mdata[['budget','domestic','international',"worldwide","mpaa","pro-loss"]]
y=mdata['roi']
```

```
RandomForestRegressor(random_state=43)
```

```
y_pred_r = rfr.predict(x_test)
```

```
from sklearn.metrics import mean_absolute_error, mean_squared_error
mae_r = mean_absolute_error(y_test, y_pred_r)
mse_r = mean_squared_error(y_test, y_pred_r)

print(f"mae_r: {mae_r}, \nmse_r: {mse_r}")
```

```
mae_r: 3366025.9361290326,
mse_r: 70526374264336.25
```

```
# Changing the parameters of the randomforest
```

```
from sklearn.ensemble import RandomForestRegressor
RFReg = RandomForestRegressor(n_estimators =10, random_state=0)
RFReg.fit(x_train, y_train)
```

```
RandomForestRegressor(n_estimators=10, random_state=0)
```

```
y_predict_rfr = RFReg.predict((x_test))

from sklearn import metrics
r_square = metrics.r2_score(y_test, y_predict_rfr)
print('R-square Error associated with Random Forest Regression is:', r_square)
```

```
R-square Error associated with Random Forest Regression is: 0.9926642002880407
```

Accuracy - 99%

Linear Regression Model

```
Out[87]: LinearRegression()

In [88]: y_pred = model.predict(x_test)

In [89]: print('Coefficients:', model.coef_)
         print('Intercept:', model.intercept_)
         print('Mean Squared errpr (MSE): %.2f'
               % mean_squared_error(y_test, y_pred))
         print('cofficient of determination (R^2): %.2f'
              % r2_score(y_test, y_pred))

         Coefficients: [-1.00000000e+00 -1.61622020e-08 -1.61622031e-08  1.00000002e+00
          -1.35973819e-09  4.09486478e-09]
         Intercept: -4.470348358154297e-08
         Mean Squared errpr (MSE): 0.00
         cofficient of determination (R^2): 1.00

In [90]: from sklearn import metrics
         print(metrics.mean_absolute_error(y_test,y_pred))
         print(metrics.mean_squared_error(y_test, y_pred))

         4.521444357270675e-08
         5.167433379444047e-15
```

The coefficient of determination is 1

And the mean square error is 0

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. The mean squared error is also known as the mean squared deviation (MSD)

The mean square is negligible

Summary of Results

- Making Movies and TV Shows contain a lot of factors for success and failure. The approach of the Creative Department and Production Department about the 'movie-making process' is totally different.

- The success of a Movie depends upon Director, Genre, Cast and Budget equally. If Studios does an analysis of past records of hit/flop movies on the box office, Worldwide collection, Return of Investment will help them to predict the future of the current projects in their hand.

- So, They can make decisions from the report of analysis, whether they want to go on the floor with that particular project or not or If they go ahead with the project they can change particular factors like Genre, Director, Cast and budget. They can clearly see the Box office Revenue performance of Particular actors, directors, cinematographers. So, they can deal with the creative department with respect to remuneration.

- Overall, After analyzing the dataset STUDIOS can have the FORMULA of Most Successful/Profitable Movie.

Thank You!