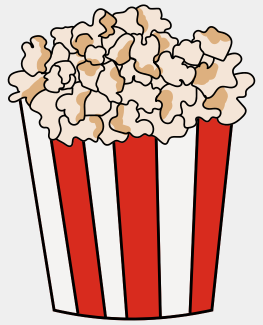
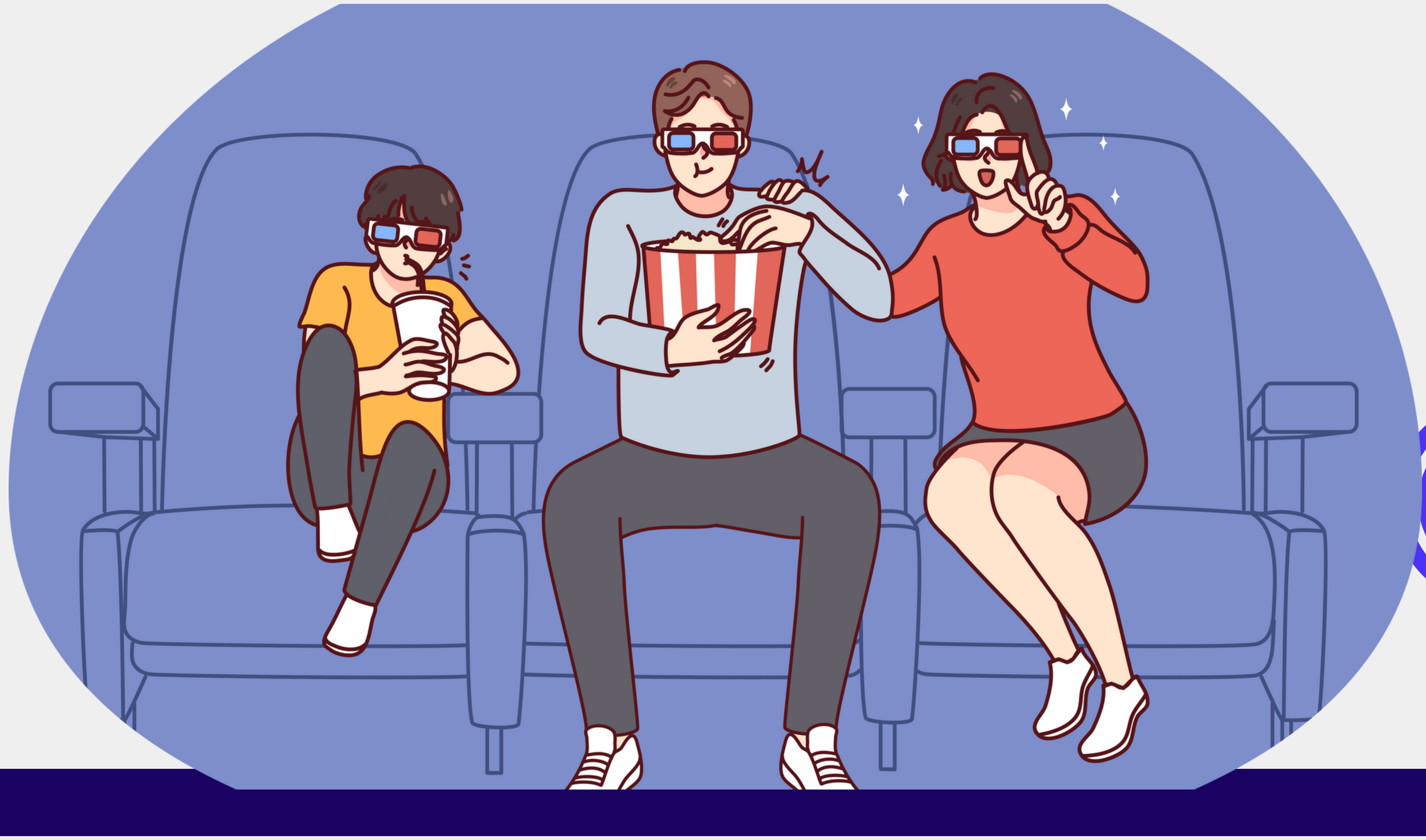
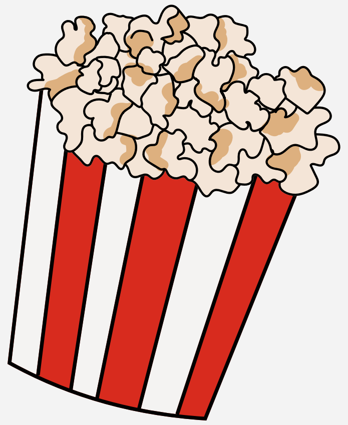
Movie Correlation





Data Used

- Movies.csv

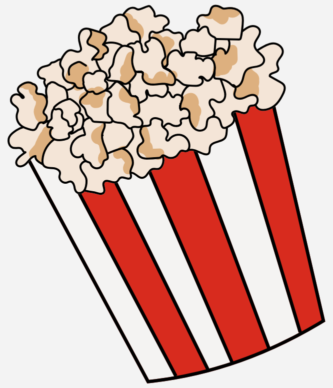


Language

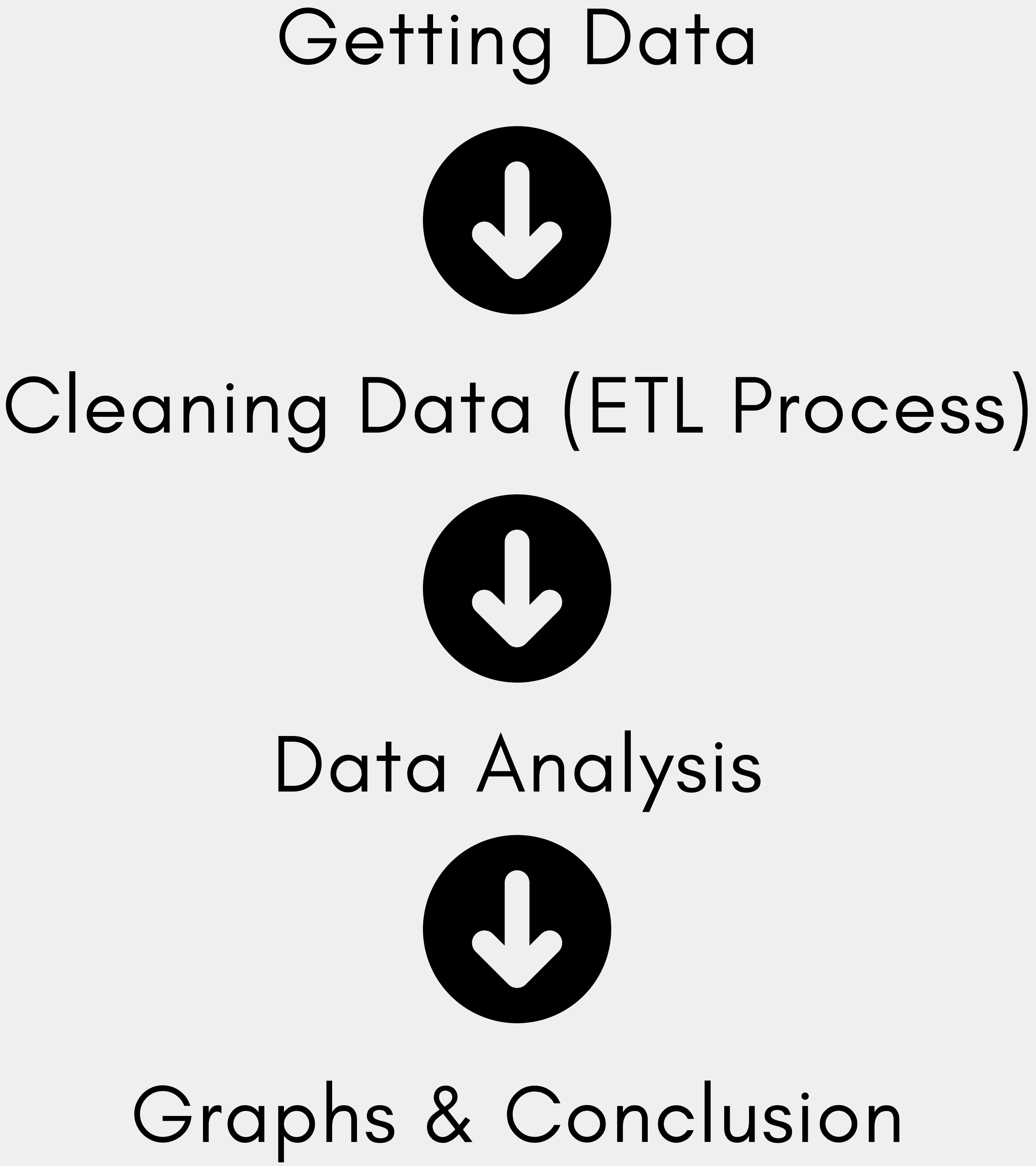
- Python

Tool

- Jupyter Notebook



Process Used In Project



Getting Data

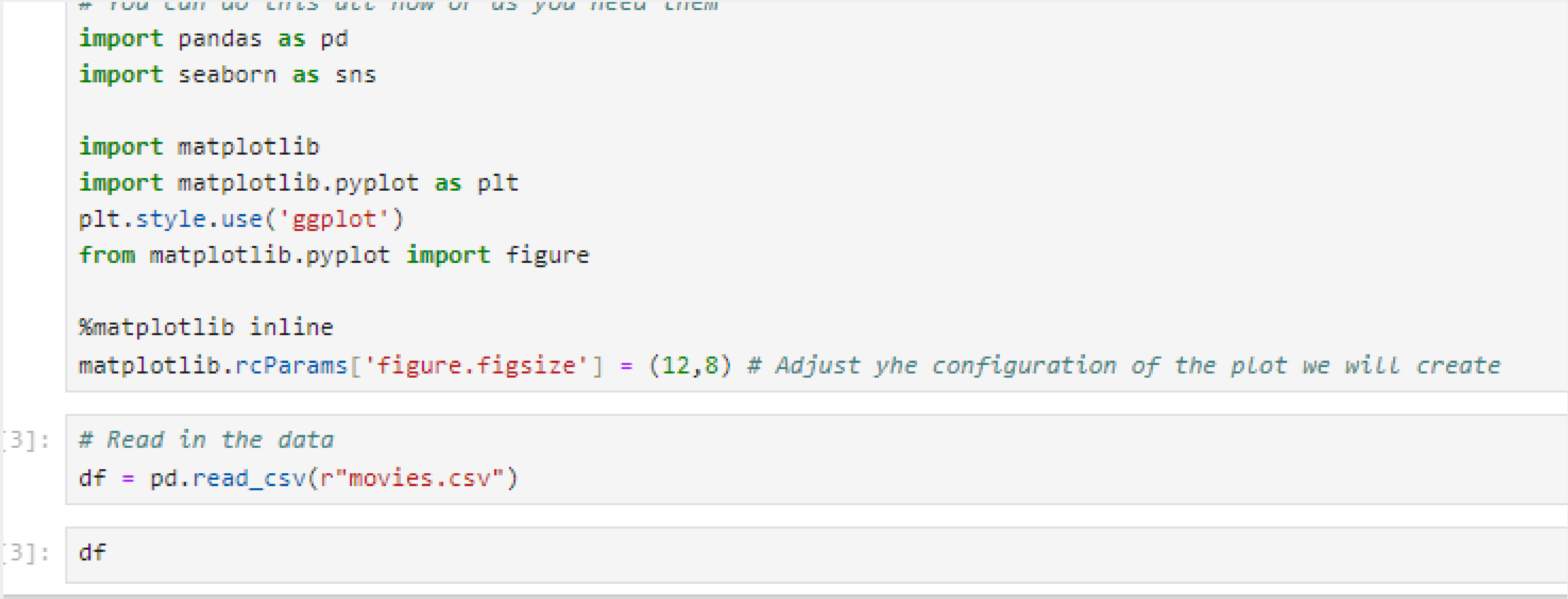
Cleaning Data (ETL Process)

Data Analysis

Graphs & Conclusion

Geting Data

Importing all the modules and reading the data



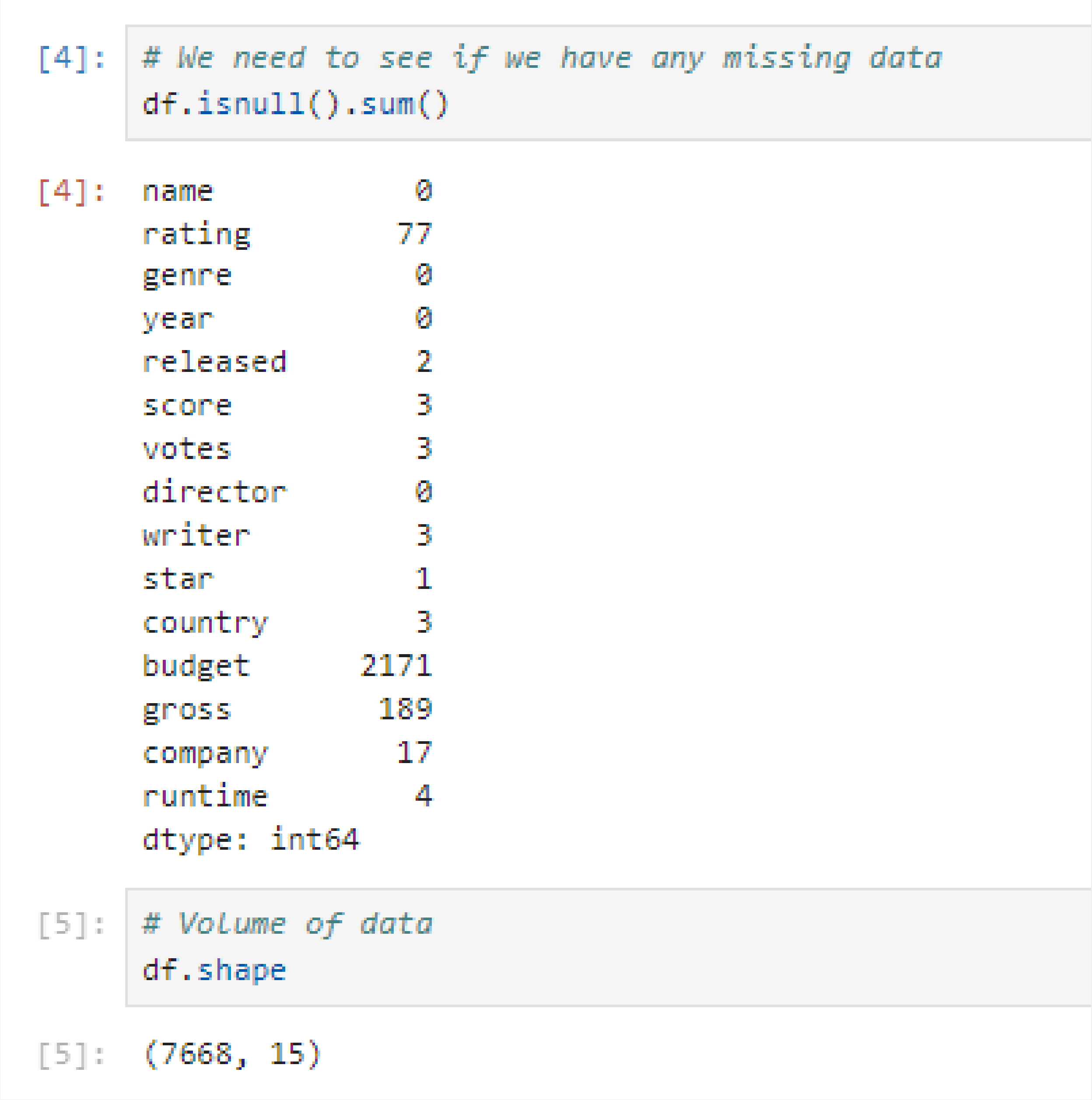
```
# You can do this all now or as you need them
import pandas as pd
import seaborn as sns

import matplotlib
import matplotlib.pyplot as plt
plt.style.use('ggplot')
from matplotlib.pyplot import figure

%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8) # Adjust yhe configuration of the plot we will create
```

```
# Read in the data
df = pd.read_csv(r"movies.csv")
```

```
df
```

Data Cleaning (ETL Process)

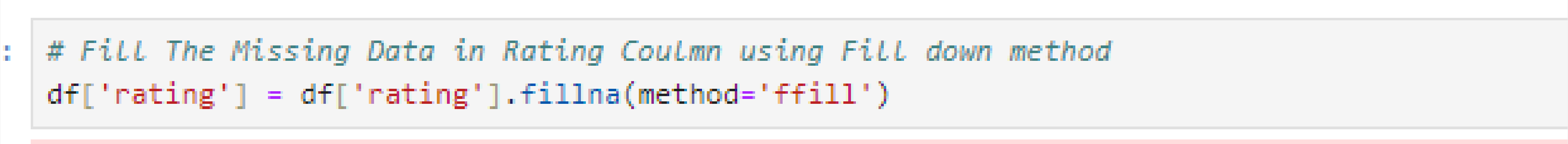
Checking the Total Number of rows and columns & Null Value



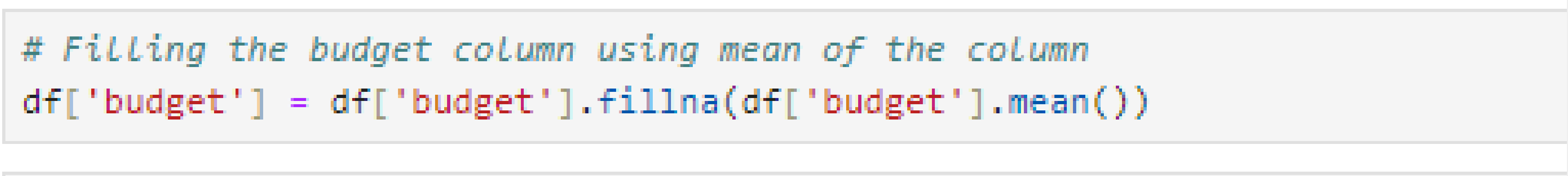
```
[4]: # We need to see if we have any missing data
     df.isnull().sum()

[4]: name           0
     rating        77
     genre          0
     year           0
     released       2
     score          3
     votes          3
     director       0
     writer         3
     star           1
     country        3
     budget      2171
     gross        189
     company       17
     runtime        4
     dtype: int64

[5]: # Volume of data
     df.shape

[5]: (7668, 15)
```

Data Cleaning (ETL Process)

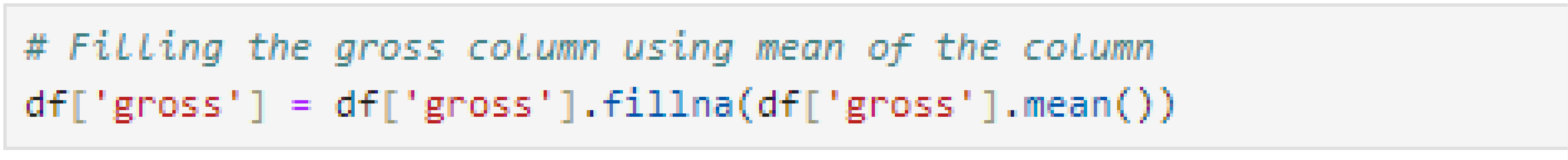
Filling the Null Value of the columns



```python
# Fill The Missing Data in Rating Coulmn using Fill down method
df['rating'] = df['rating'].fillna(method='ffill')
```

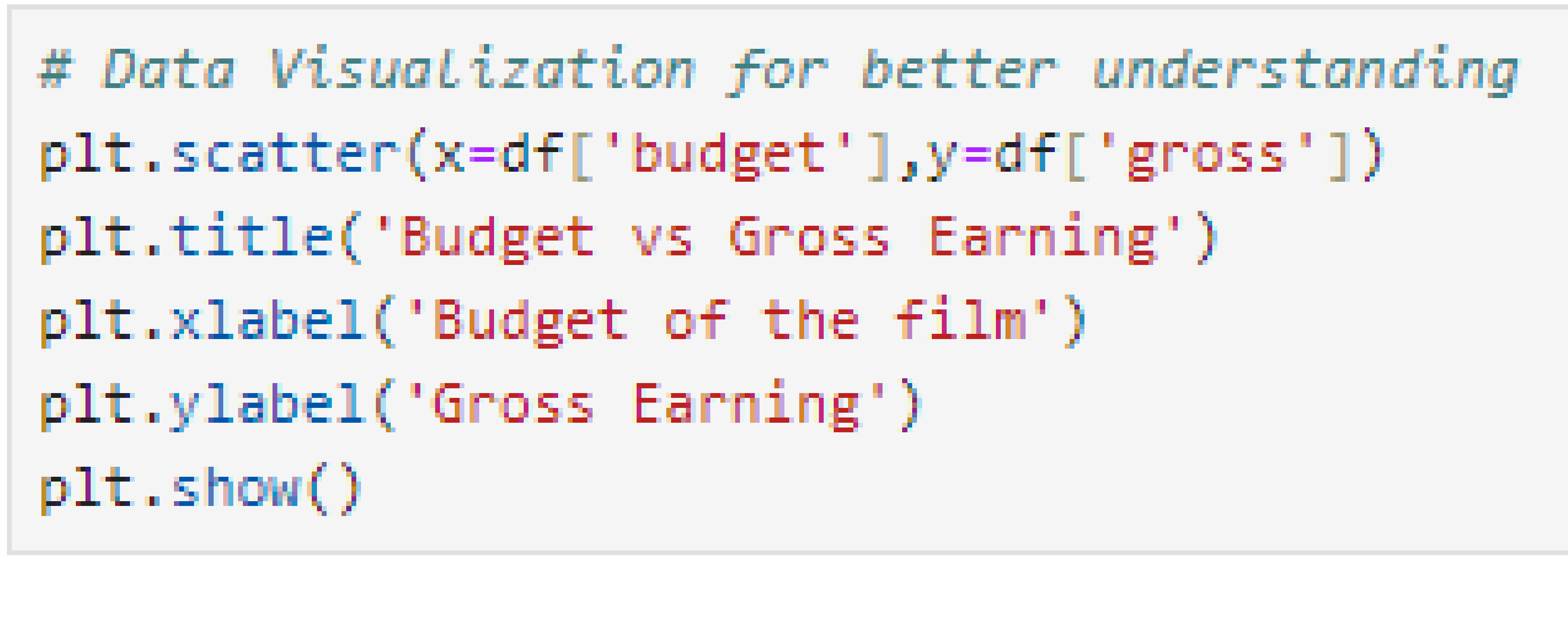


```python
# Filling the budget column using mean of the column
df['budget'] = df['budget'].fillna(df['budget'].mean())
```



```python
# Filling the gross column using mean of the column
df['gross'] = df['gross'].fillna(df['gross'].mean())
```

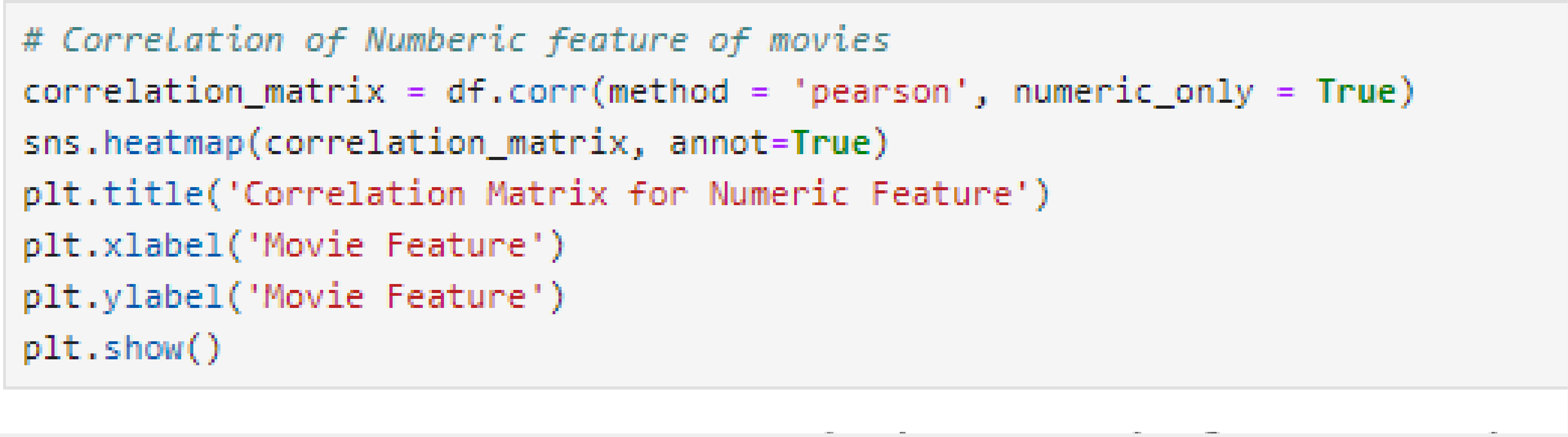
Data Analysis



```
# Data Visualization for better understanding
plt.scatter(x=df['budget'],y=df['gross'])
plt.title('Budget vs Gross Earning')
plt.xlabel('Budget of the film')
plt.ylabel('Gross Earning')
plt.show()
```

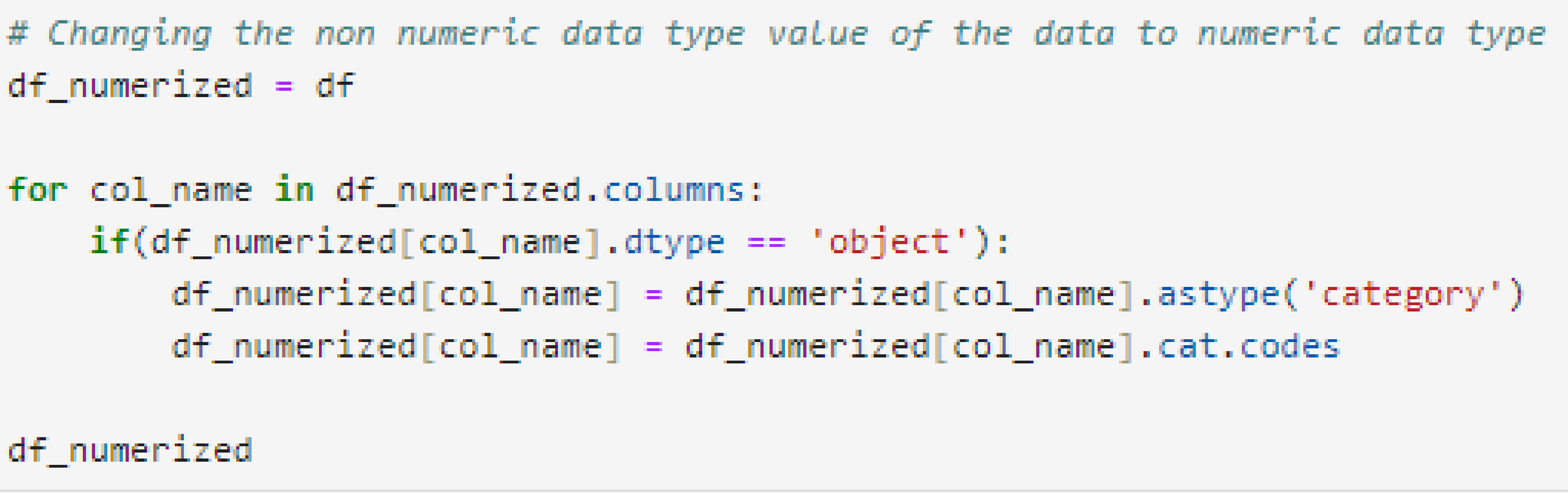
Finding Relation between budget of the movie and the gross earning of the movie

Correlation between the numeric feature of the data



```
# Correlation of Numberic feature of movies
correlation_matrix = df.corr(method = 'pearson', numeric_only = True)
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matrix for Numeric Feature')
plt.xlabel('Movie Feature')
plt.ylabel('Movie Feature')
plt.show()
```

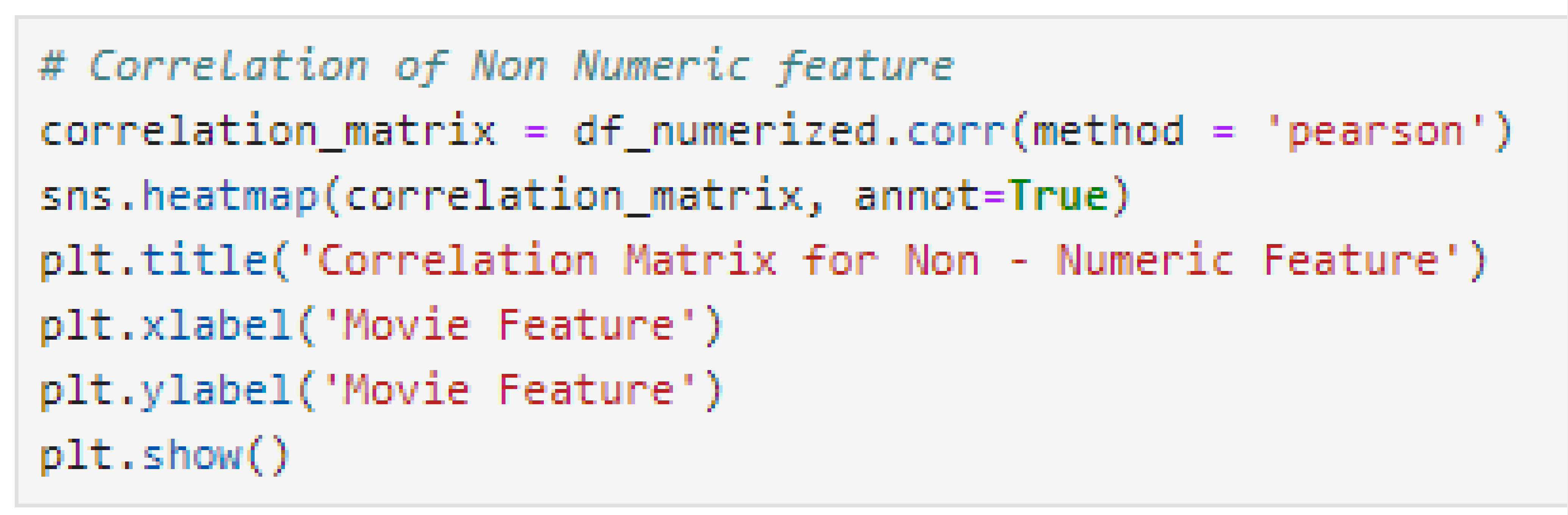
Data Analysis



```
# Changing the non numeric data type value of the data to numeric data type
df_numerized = df

for col_name in df_numerized.columns:
    if(df_numerized[col_name].dtype == 'object'):
        df_numerized[col_name] = df_numerized[col_name].astype('category')
        df_numerized[col_name] = df_numerized[col_name].cat.codes

df_numerized
```

Changing the non numeric value data type of the data to the numeric value data type of the data

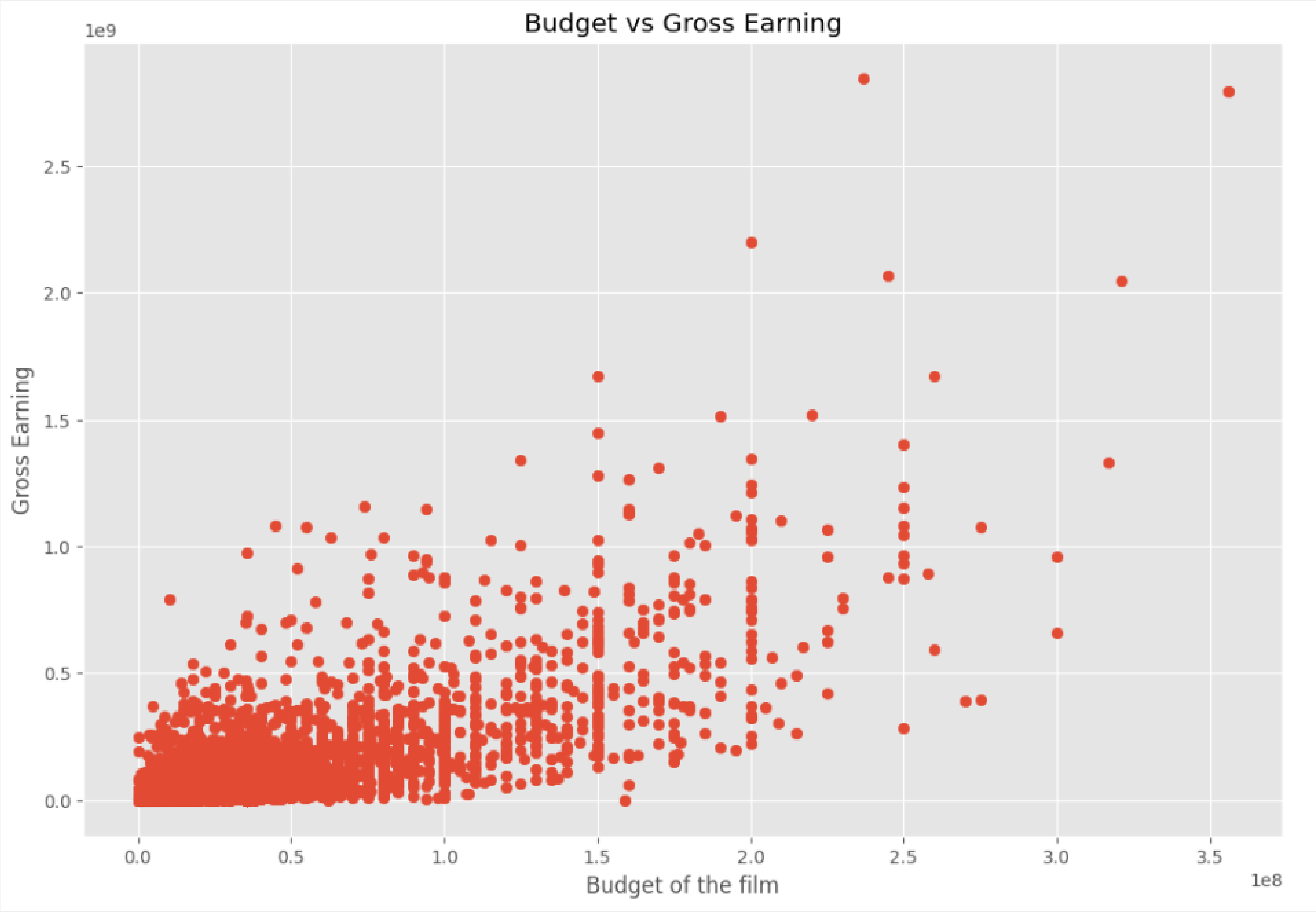
Correlation between the non numeric feature of the data



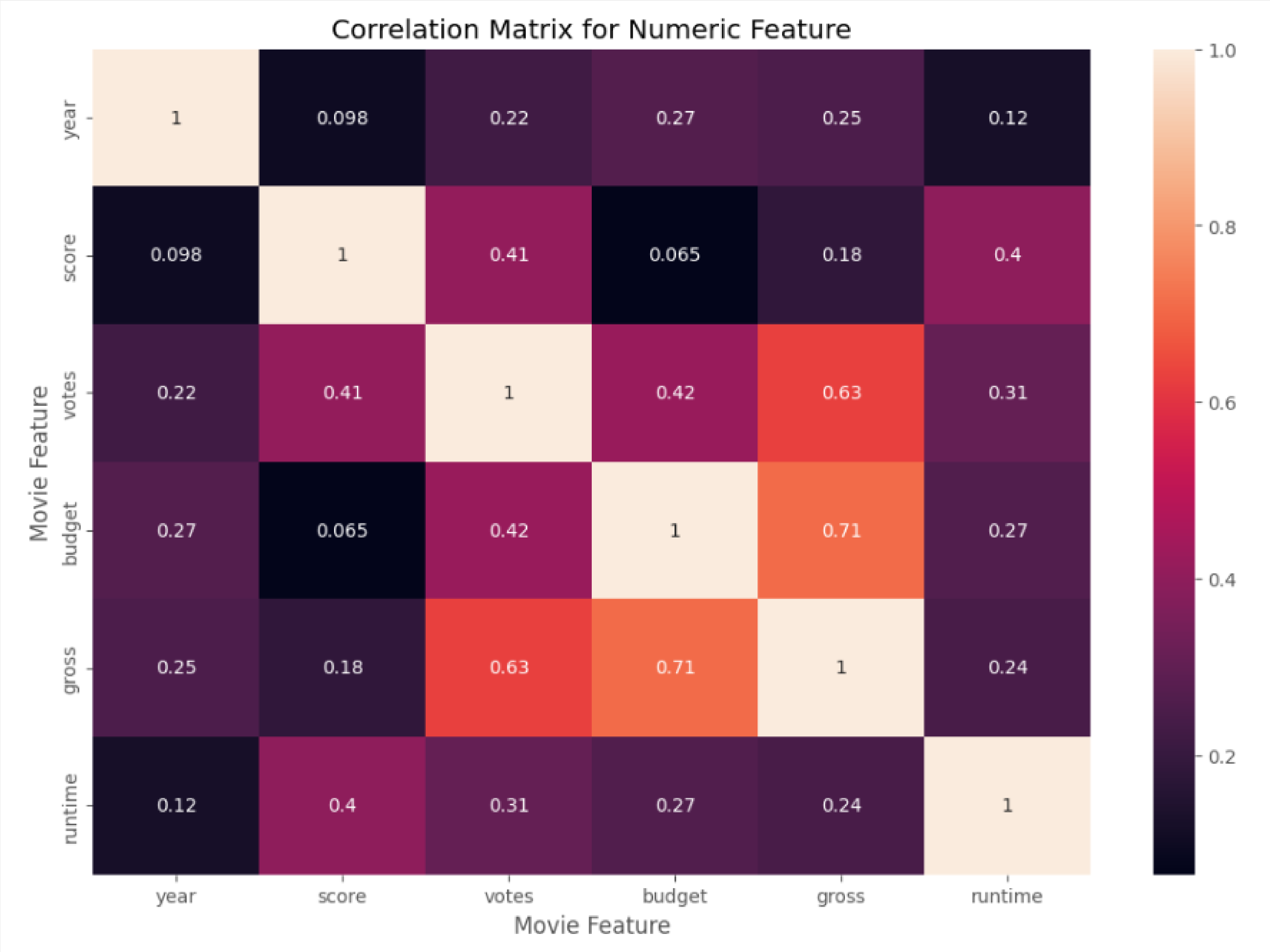
```
# Correlation of Non Numeric feature
correlation_matrix = df_numerized.corr(method = 'pearson')
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matrix for Non - Numeric Feature')
plt.xlabel('Movie Feature')
plt.ylabel('Movie Feature')
plt.show()
```

Graphs

Scatter plot shows relationship between Budget of the movie and the Gross earning of the movie



Graphs



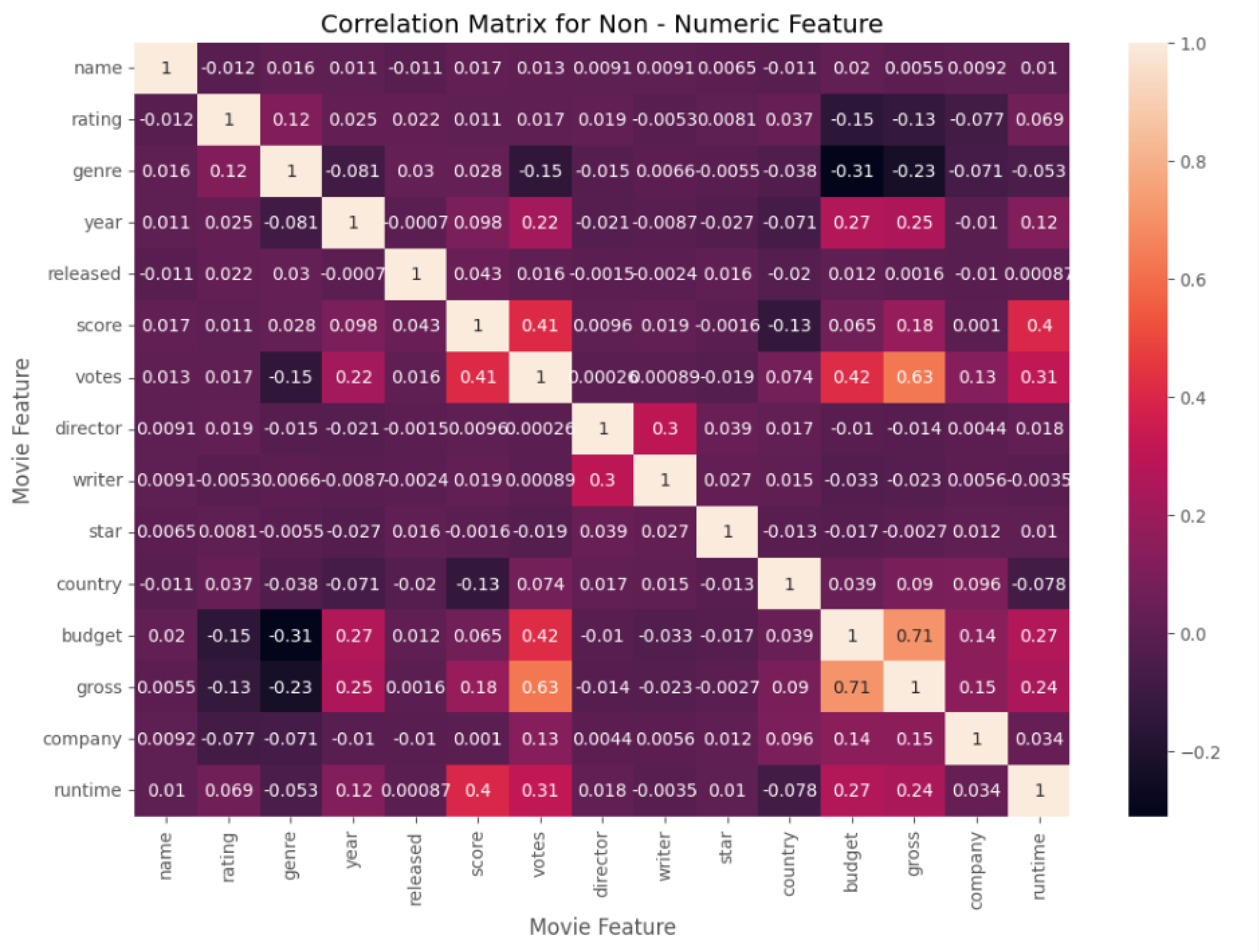
Correlation Matrix for Numeric Feature

	year	score	votes	budget	gross	runtime
year	1	0.098	0.22	0.27	0.25	0.12
score	0.098	1	0.41	0.065	0.18	0.4
votes	0.22	0.41	1	0.42	0.63	0.31
budget	0.27	0.065	0.42	1	0.71	0.27
gross	0.25	0.18	0.63	0.71	1	0.24
runtime	0.12	0.4	0.31	0.27	0.24	1

The Heatmap here shows the correlation between the numeric feature of the movie

Graphs

The Heatmap here shows the correlation between the non numeric feature of the movie



Correlation Matrix for Non - Numeric Feature

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
name	1	-0.012	0.016	0.011	-0.011	0.017	0.013	0.0091	0.0091	0.0065	-0.011	0.02	0.0055	0.0092	0.01
rating	-0.012	1	0.12	0.025	0.022	0.011	0.017	0.019	-0.0053	0.0081	0.037	-0.15	-0.13	-0.077	0.069
genre	0.016	0.12	1	-0.081	0.03	0.028	-0.15	-0.015	0.0066	-0.0055	-0.038	-0.31	-0.23	-0.071	-0.053
year	0.011	0.025	-0.081	1	-0.0007	0.098	0.22	-0.021	-0.0087	-0.027	-0.071	0.27	0.25	-0.01	0.12
released	-0.011	0.022	0.03	-0.0007	1	0.043	0.016	-0.0015	-0.0024	0.016	-0.02	0.012	0.0016	-0.01	0.00087
score	0.017	0.011	0.028	0.098	0.043	1	0.41	0.0096	0.019	-0.0016	-0.13	0.065	0.18	0.001	0.4
votes	0.013	0.017	-0.15	0.22	0.016	0.41	1	0.00026	0.00089	-0.019	0.074	0.42	0.63	0.13	0.31
director	0.0091	0.019	-0.015	-0.021	-0.0015	0.0096	0.00026	1	0.3	0.039	0.017	-0.01	-0.014	0.0044	0.018
writer	0.0091	-0.0053	0.0066	-0.0087	-0.0024	0.019	0.00089	0.3	1	0.027	0.015	-0.033	-0.023	0.0056	-0.0035
star	0.0065	0.0081	-0.0055	-0.027	0.016	-0.0016	-0.019	0.039	0.027	1	-0.013	-0.017	-0.0027	0.012	0.01
country	-0.011	0.037	-0.038	-0.071	-0.02	-0.13	0.074	0.017	0.015	-0.013	1	0.039	0.09	0.096	-0.078
budget	0.02	-0.15	-0.31	0.27	0.012	0.065	0.42	-0.01	-0.033	-0.017	0.039	1	0.71	0.14	0.27
gross	0.0055	-0.13	-0.23	0.25	0.0016	0.18	0.63	-0.014	-0.023	-0.0027	0.09	0.71	1	0.15	0.24
company	0.0092	-0.077	-0.071	-0.01	-0.01	0.001	0.13	0.0044	0.0056	0.012	0.096	0.14	0.15	1	0.034
runtime	0.01	0.069	-0.053	0.12	0.00087	0.4	0.31	0.018	-0.0035	0.01	-0.078	0.27	0.24	0.034	1

Conclusion

The Conclusion of the project is that the Votes and Gross Earning of the movies has the highest correlation.

Thank you for your time.

CONTACT

Feel free to reach out:

Email: tiwariv617@gmail.com
LinkedIn: www.linkedin.com/in/vishal-tiwari-6a380a2aa
Portifolio: https://vishaltiwari1.github.io/Portfolio/