# EDS MINI PROJECT

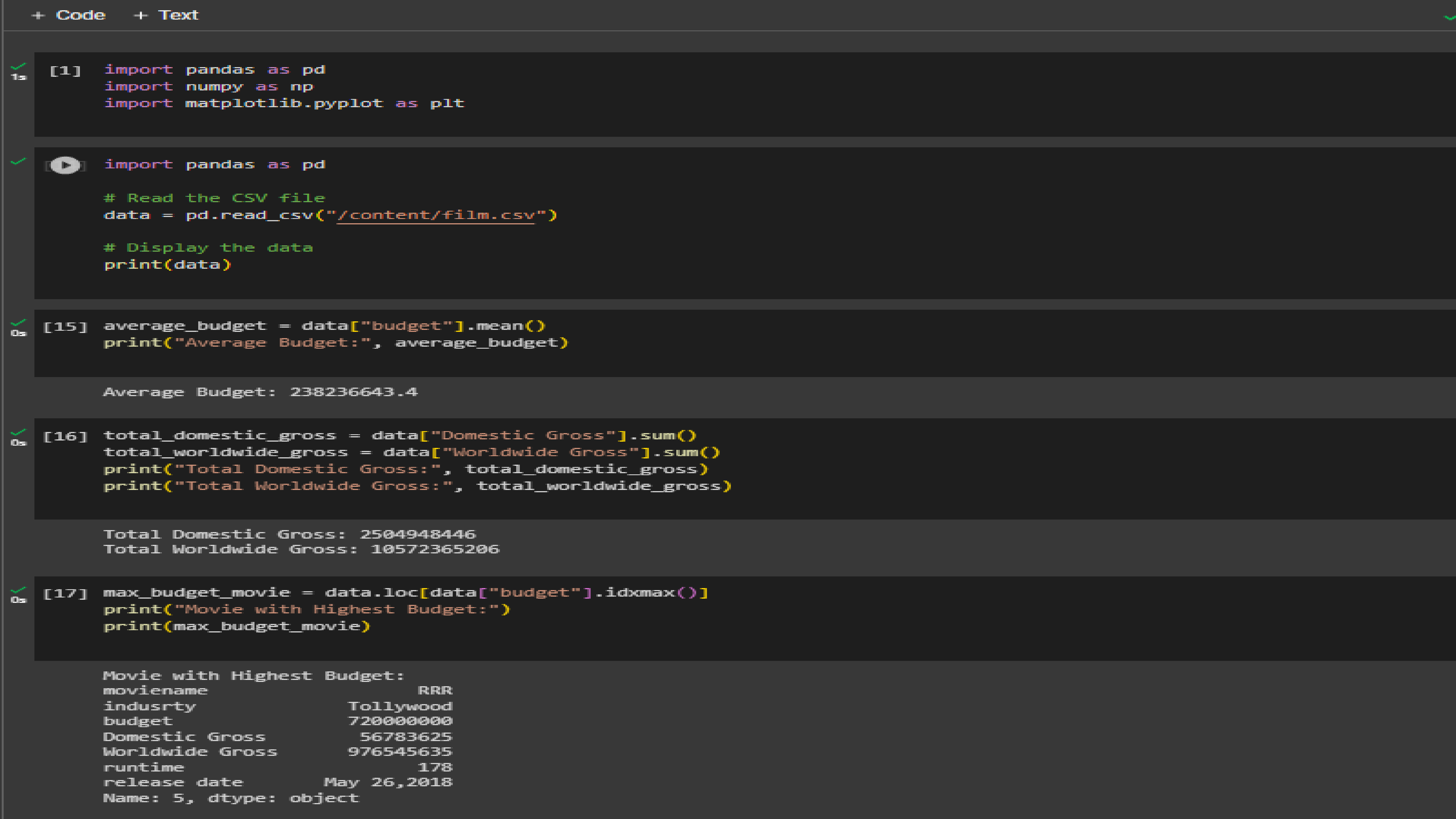
Name : Siddharth Dnyaneshwar Tambe

Roll No. : 364

PRN : 202201090172

Div : C

Batch : C4



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
import pandas as pd

# Read the CSV file
data = pd.read_csv("/content/film.csv")

# Display the data
print(data)
```

```
average_budget = data["budget"].mean()
print("Average Budget:", average_budget)
```

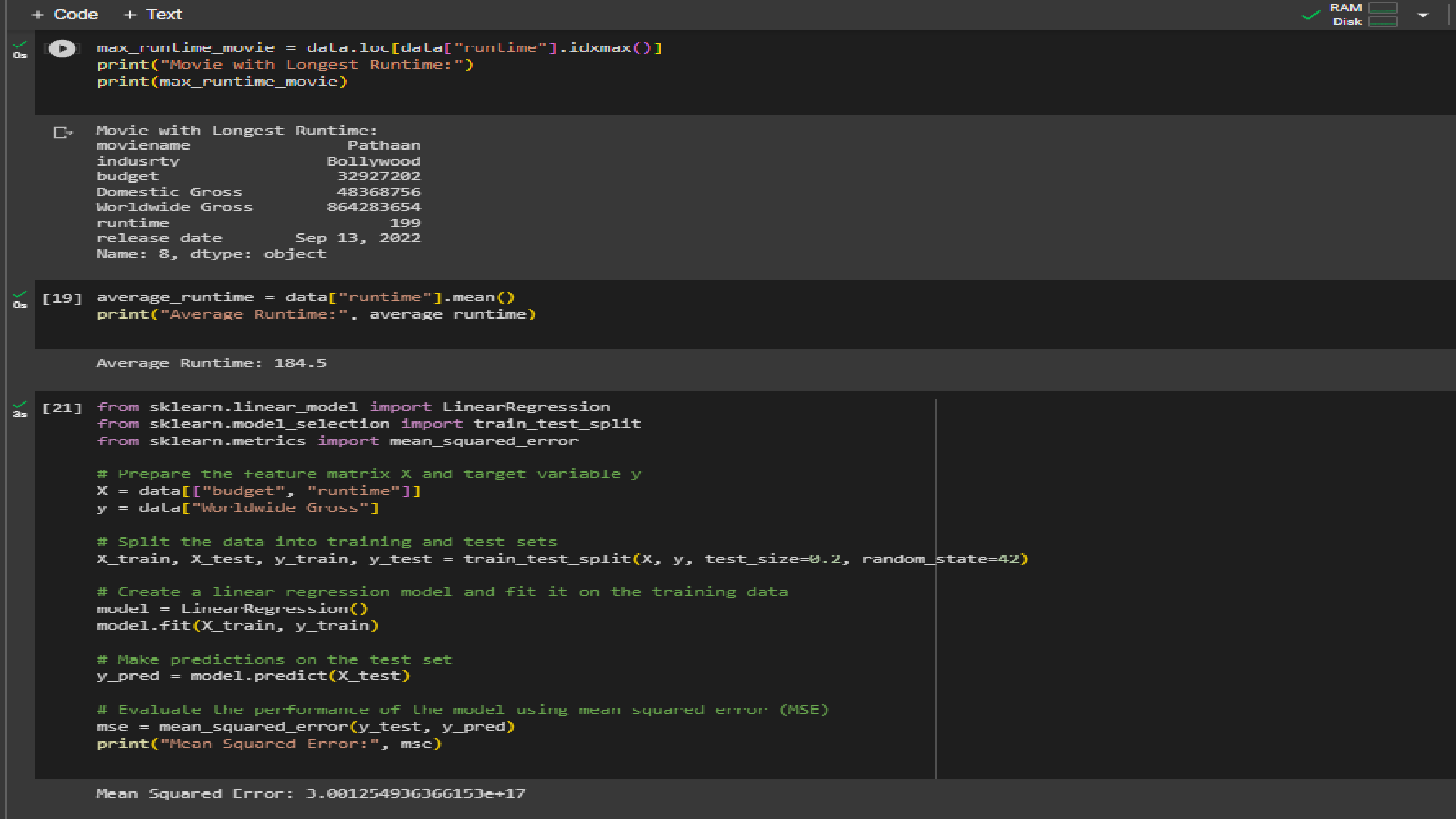
```
Average Budget: 238236643.4
```

```
total_domestic_gross = data["Domestic Gross"].sum()
total_worldwide_gross = data["Worldwide Gross"].sum()
print("Total Domestic Gross:", total_domestic_gross)
print("Total Worldwide Gross:", total_worldwide_gross)
```

```
Total Domestic Gross: 2504948446
Total Worldwide Gross: 10572365206
```

```
max_budget_movie = data.loc[data["budget"].idxmax()]
print("Movie with Highest Budget:")
print(max_budget_movie)
```

```
Movie with Highest Budget:
moviename                      RRR
indusrty                 Tollywood
budget                   720000000
Domestic Gross            56783625
Worldwide Gross          976545635
runtime                        178
release date          May 26,2018
Name: 5, dtype: object
```



```
max_runtime_movie = data.loc[data["runtime"].idxmax()]
print("Movie with Longest Runtime:")
print(max_runtime_movie)
```

```
Movie with Longest Runtime:
moviename                  Pathaan
indusrty                 Bollywood
budget                    32927202
Domestic Gross            48368756
Worldwide Gross          864283654
runtime                        199
release date          Sep 13, 2022
Name: 8, dtype: object
```

[19]

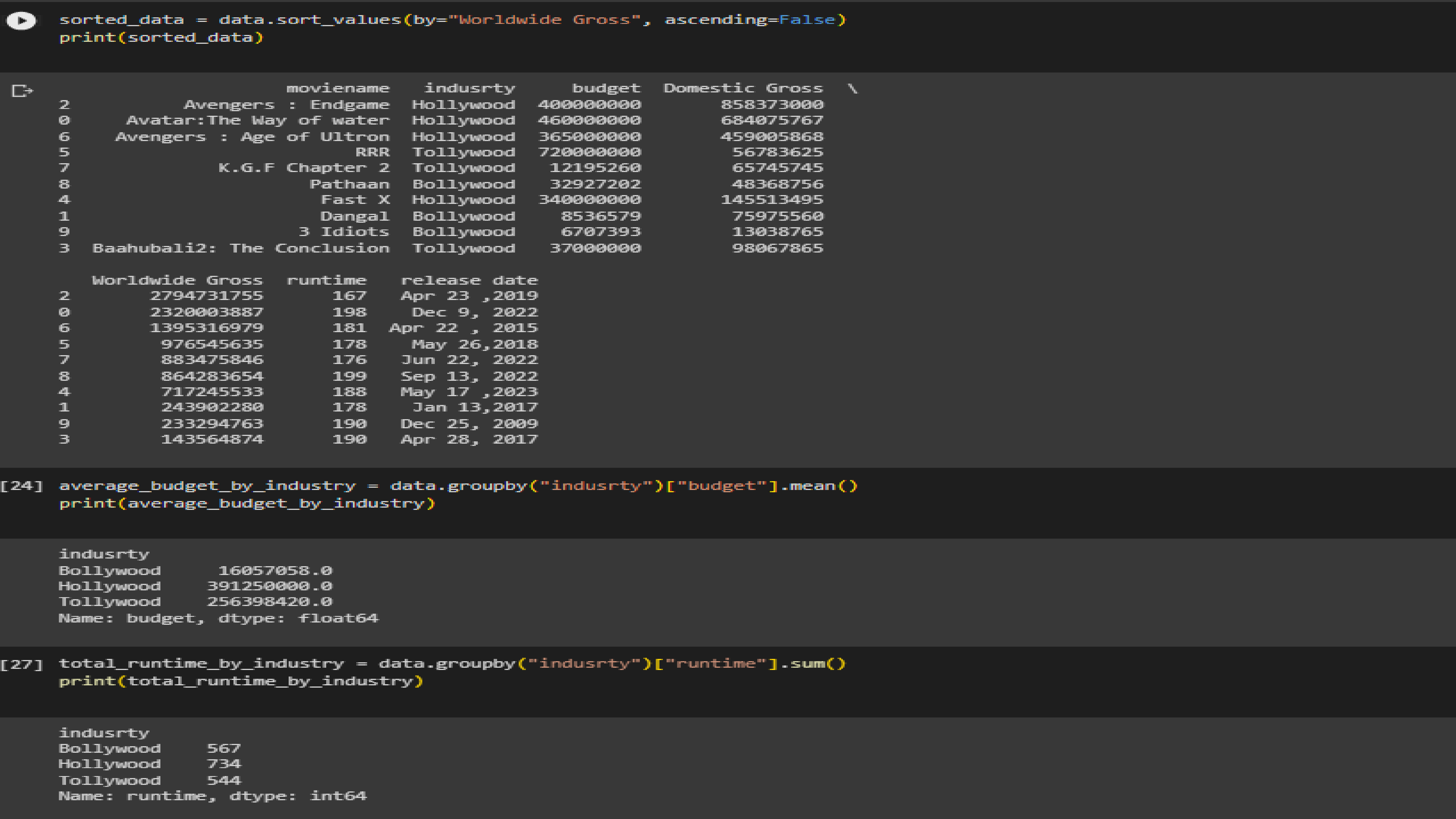
```
average_runtime = data["runtime"].mean()
print("Average Runtime:", average_runtime)
```

```
Average Runtime: 184.5
```

[21]

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

# Prepare the feature matrix X and target variable y
X = data[["budget", "runtime"]]
y = data["Worldwide Gross"]

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create a linear regression model and fit it on the training data
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the performance of the model using mean squared error (MSE)
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

```
Mean Squared Error: 3.001254936366153e+17
```



```
sorted_data = data.sort_values(by="Worldwide Gross", ascending=False)
print(sorted_data)
```

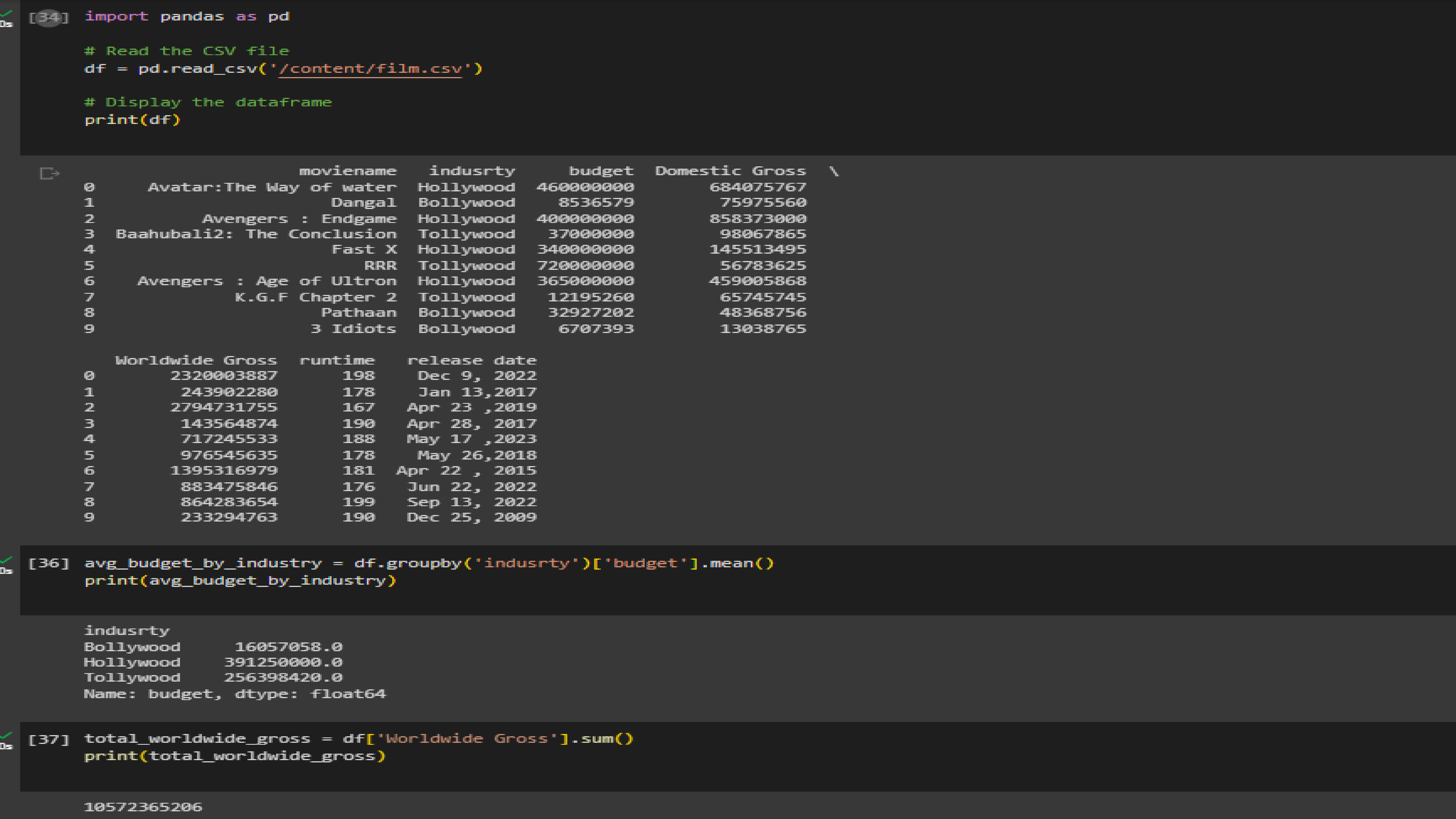
```
                    moviename   indusrty     budget  Domestic Gross  \
2           Avengers : Endgame  Hollywood  400000000       858373000
0      Avatar:The Way of water  Hollywood  460000000       684075767
6    Avengers : Age of Ultron   Hollywood  365000000       459005868
5                          RRR  Tollywood  720000000        56783625
7            K.G.F Chapter 2    Tollywood   12195260        65745745
8                      Pathaan  Bollywood   32927202        48368756
4                       Fast X  Hollywood  340000000       145513495
1                       Dangal  Bollywood    8536579        75975560
9                     3 Idiots  Bollywood    6707393        13038765
3  Baahubali2: The Conclusion   Tollywood   37000000        98067865

   Worldwide Gross  runtime    release date
2       2794731755      167    Apr 23 ,2019
0       2320003887      198     Dec 9, 2022
6       1395316979      181  Apr 22 , 2015
5        976545635      178    May 26,2018
7        883475846      176    Jun 22, 2022
8        864283654      199    Sep 13, 2022
4        717245533      188    May 17 ,2023
1        243902280      178    Jan 13,2017
9        233294763      190    Dec 25, 2009
3        143564874      190    Apr 28, 2017
```

```
[24] average_budget_by_industry = data.groupby("indusrty")["budget"].mean()
     print(average_budget_by_industry)
```

```
indusrty
Bollywood     16057058.0
Hollywood    391250000.0
Tollywood    256398420.0
Name: budget, dtype: float64
```

```
[27] total_runtime_by_industry = data.groupby("indusrty")["runtime"].sum()
     print(total_runtime_by_industry)
```

```
indusrty
Bollywood    567
Hollywood    734
Tollywood    544
Name: runtime, dtype: int64
```



[34]

```
import pandas as pd

# Read the CSV file
df = pd.read_csv('/content/film.csv')

# Display the dataframe
print(df)
```

```
                  moviename   indusrty      budget  Domestic Gross  \
0     Avatar:The Way of water  Hollywood  460000000       684075767
1                     Dangal  Bollywood    8536579        75975560
2          Avengers : Endgame  Hollywood  400000000       858373000
3  Baahubali2: The Conclusion  Tollywood   37000000        98067865
4                     Fast X  Hollywood  340000000       145513495
5                        RRR  Tollywood  720000000        56783625
6    Avengers : Age of Ultron  Hollywood  365000000       459005868
7            K.G.F Chapter 2  Tollywood   12195260        65745745
8                    Pathaan  Bollywood   32927202        48368756
9                   3 Idiots  Bollywood    6707393        13038765

   Worldwide Gross  runtime   release date
0       2320003887      198    Dec 9, 2022
1        243902280      178    Jan 13,2017
2       2794731755      167   Apr 23 ,2019
3        143564874      190   Apr 28, 2017
4        717245533      188   May 17 ,2023
5        976545635      178    May 26,2018
6       1395316979      181  Apr 22 , 2015
7        883475846      176   Jun 22, 2022
8        864283654      199   Sep 13, 2022
9        233294763      190   Dec 25, 2009
```

[36]

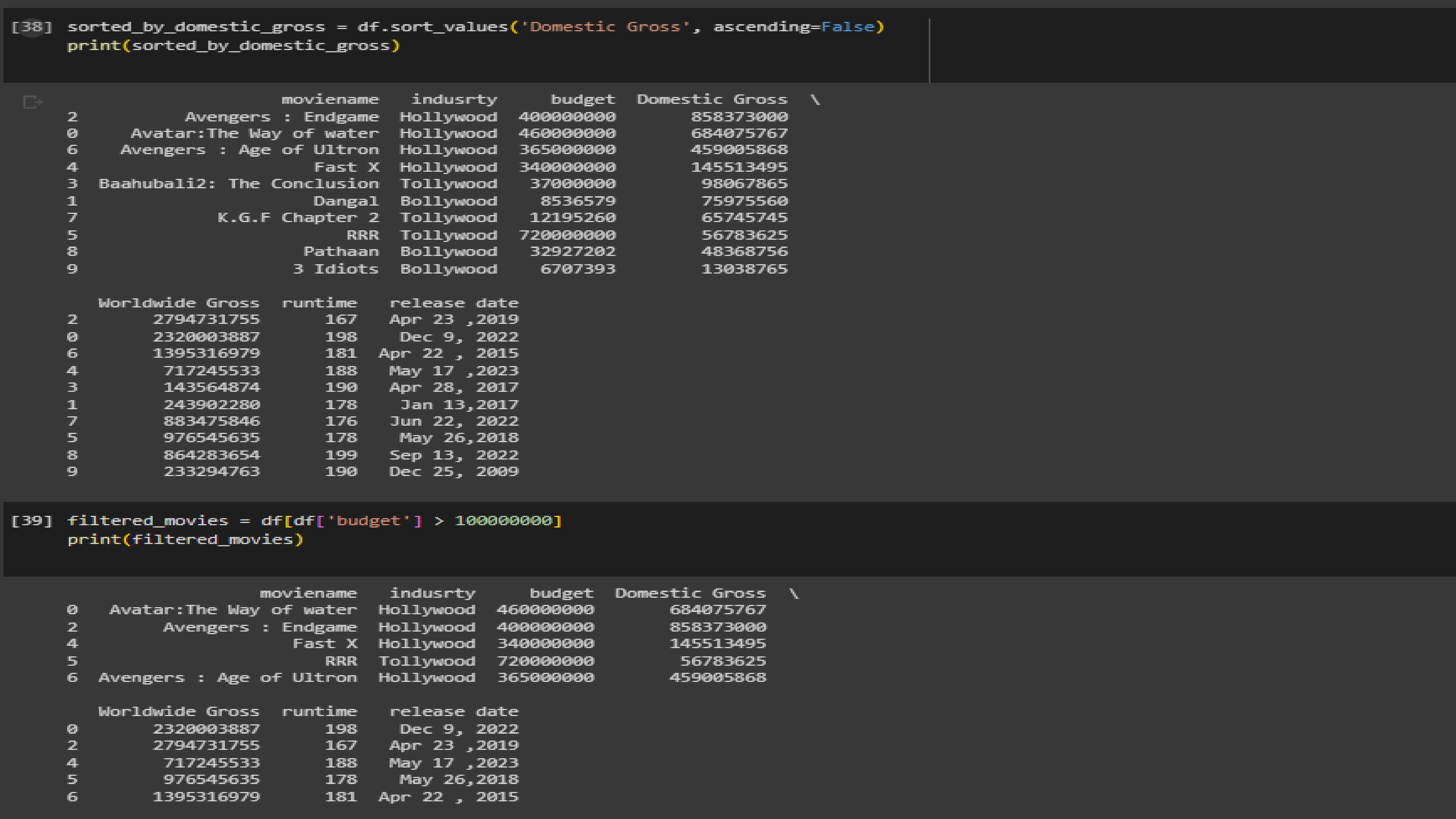
```
avg_budget_by_industry = df.groupby('indusrty')['budget'].mean()
print(avg_budget_by_industry)
```

```
indusrty
Bollywood     16057058.0
Hollywood    391250000.0
Tollywood    256398420.0
Name: budget, dtype: float64
```

[37]

```
total_worldwide_gross = df['Worldwide Gross'].sum()
print(total_worldwide_gross)
```

```
10572365206
```

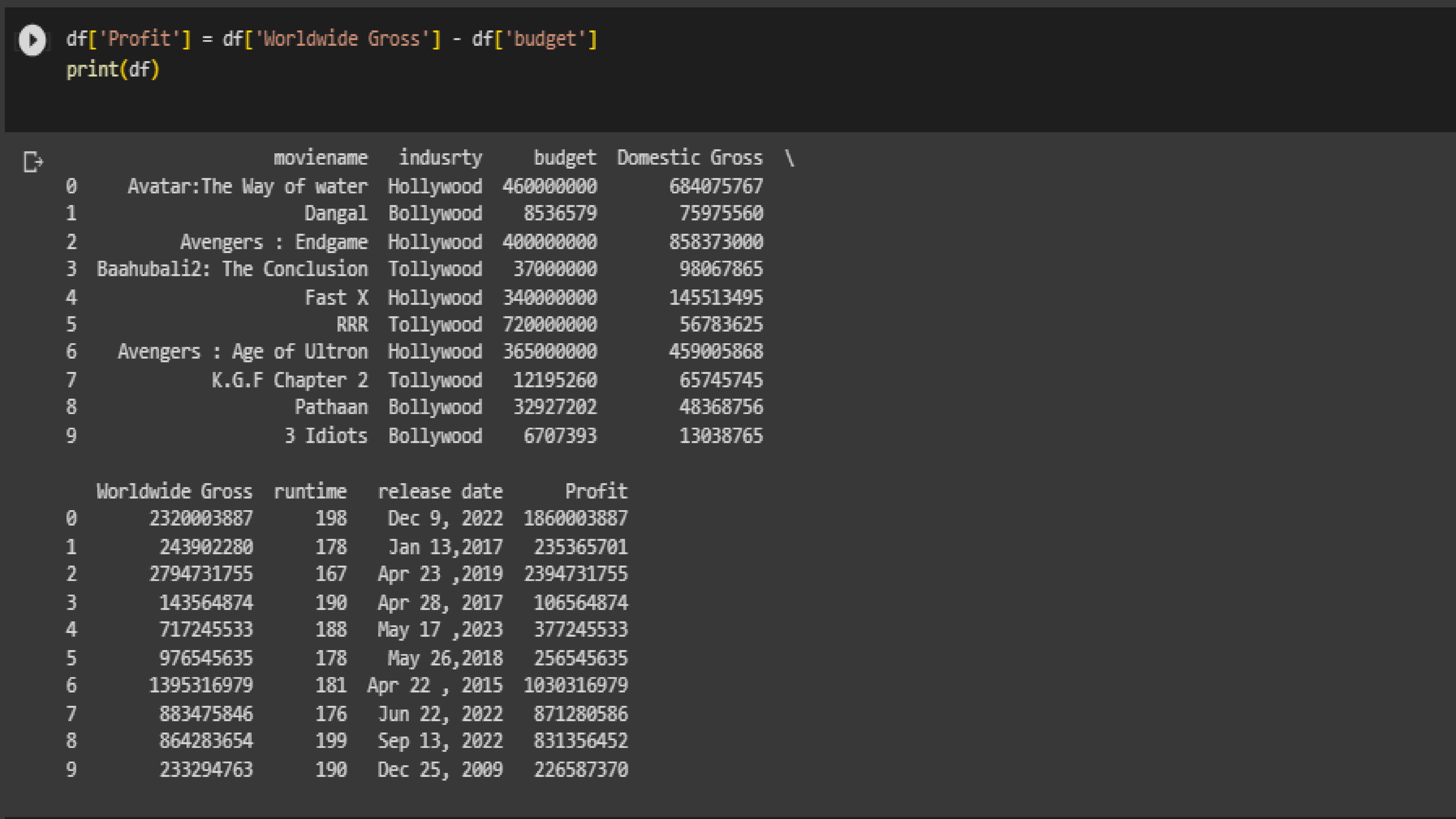


```
[38] sorted_by_domestic_gross = df.sort_values('Domestic Gross', ascending=False)
     print(sorted_by_domestic_gross)
```

```
                     moviename   indusrty     budget  Domestic Gross  \
2           Avengers : Endgame  Hollywood  400000000       858373000
0      Avatar:The Way of water  Hollywood  460000000       684075767
6    Avengers : Age of Ultron  Hollywood  365000000       459005868
4                      Fast X  Hollywood  340000000       145513495
3  Baahubali2: The Conclusion  Tollywood   37000000        98067865
1                      Dangal  Bollywood    8536579        75975560
7          K.G.F Chapter 2  Tollywood   12195260        65745745
5                          RRR  Tollywood  720000000        56783625
8                      Pathaan  Bollywood   32927202        48368756
9                     3 Idiots  Bollywood    6707393        13038765

   Worldwide Gross  runtime   release date
2       2794731755      167   Apr 23 ,2019
0       2320003887      198    Dec 9, 2022
6       1395316979      181  Apr 22 , 2015
4        717245533      188   May 17 ,2023
3        143564874      190   Apr 28, 2017
1        243902280      178    Jan 13,2017
7        883475846      176   Jun 22, 2022
5        976545635      178    May 26,2018
8        864283654      199   Sep 13, 2022
9        233294763      190   Dec 25, 2009
```

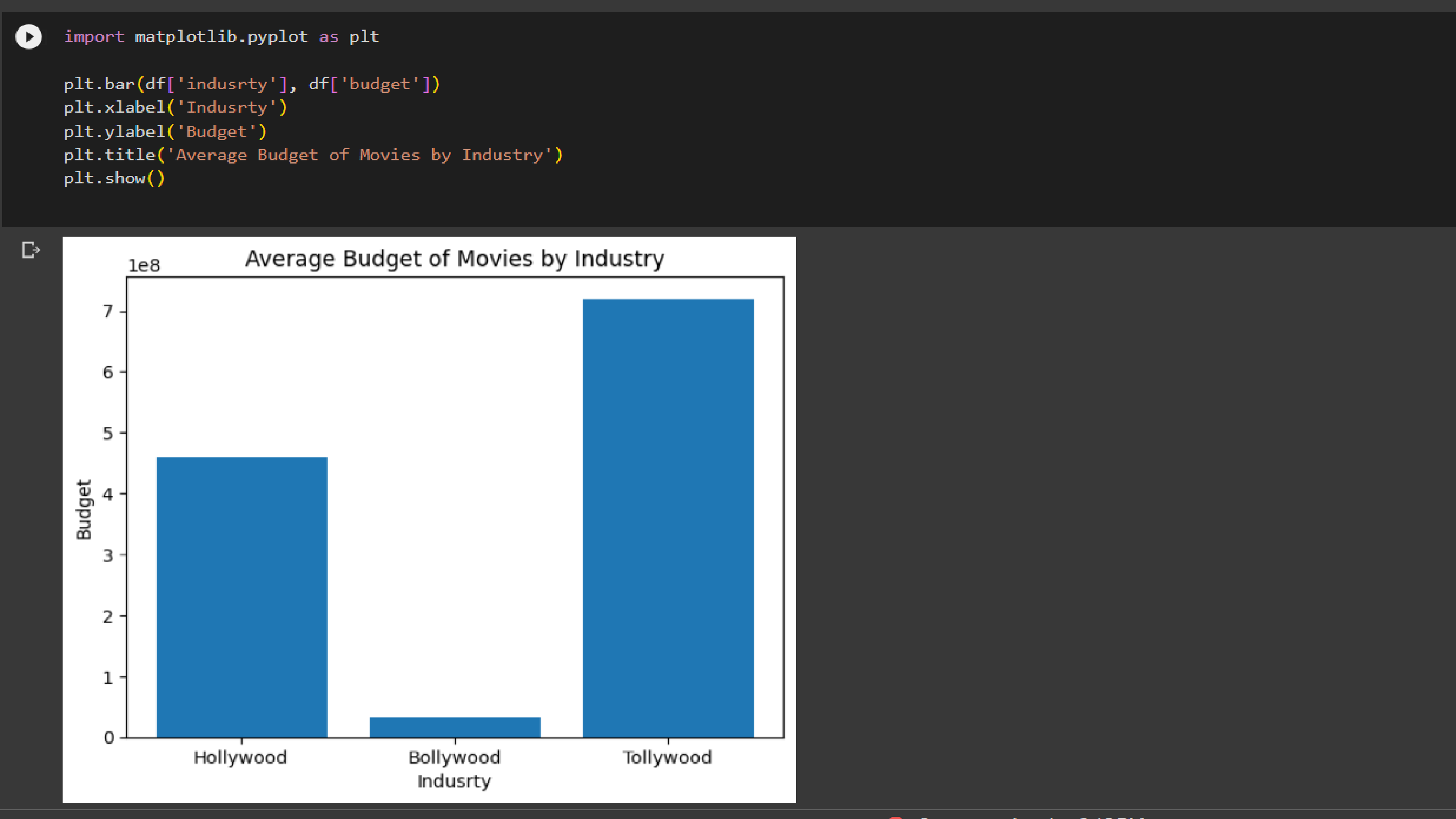
```
[39] filtered_movies = df[df['budget'] > 100000000]
     print(filtered_movies)
```

```
                  moviename   indusrty     budget  Domestic Gross  \
0   Avatar:The Way of water  Hollywood  460000000       684075767
2        Avengers : Endgame  Hollywood  400000000       858373000
4                    Fast X  Hollywood  340000000       145513495
5                       RRR  Tollywood  720000000        56783625
6  Avengers : Age of Ultron  Hollywood  365000000       459005868

   Worldwide Gross  runtime   release date
0       2320003887      198    Dec 9, 2022
2       2794731755      167   Apr 23 ,2019
4        717245533      188   May 17 ,2023
5        976545635      178    May 26,2018
6       1395316979      181  Apr 22 , 2015
```

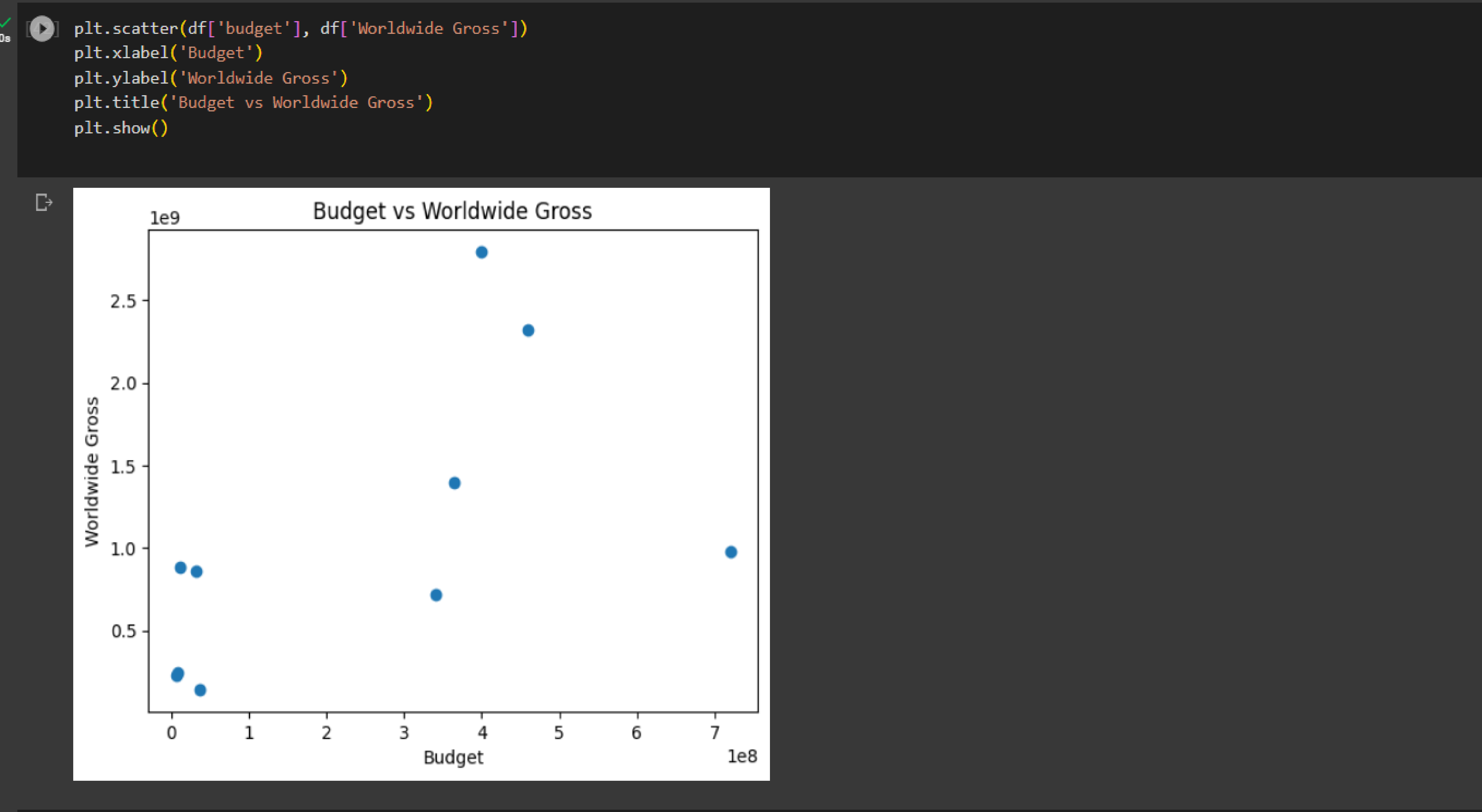


```
df['Profit'] = df['Worldwide Gross'] - df['budget']
print(df)
```

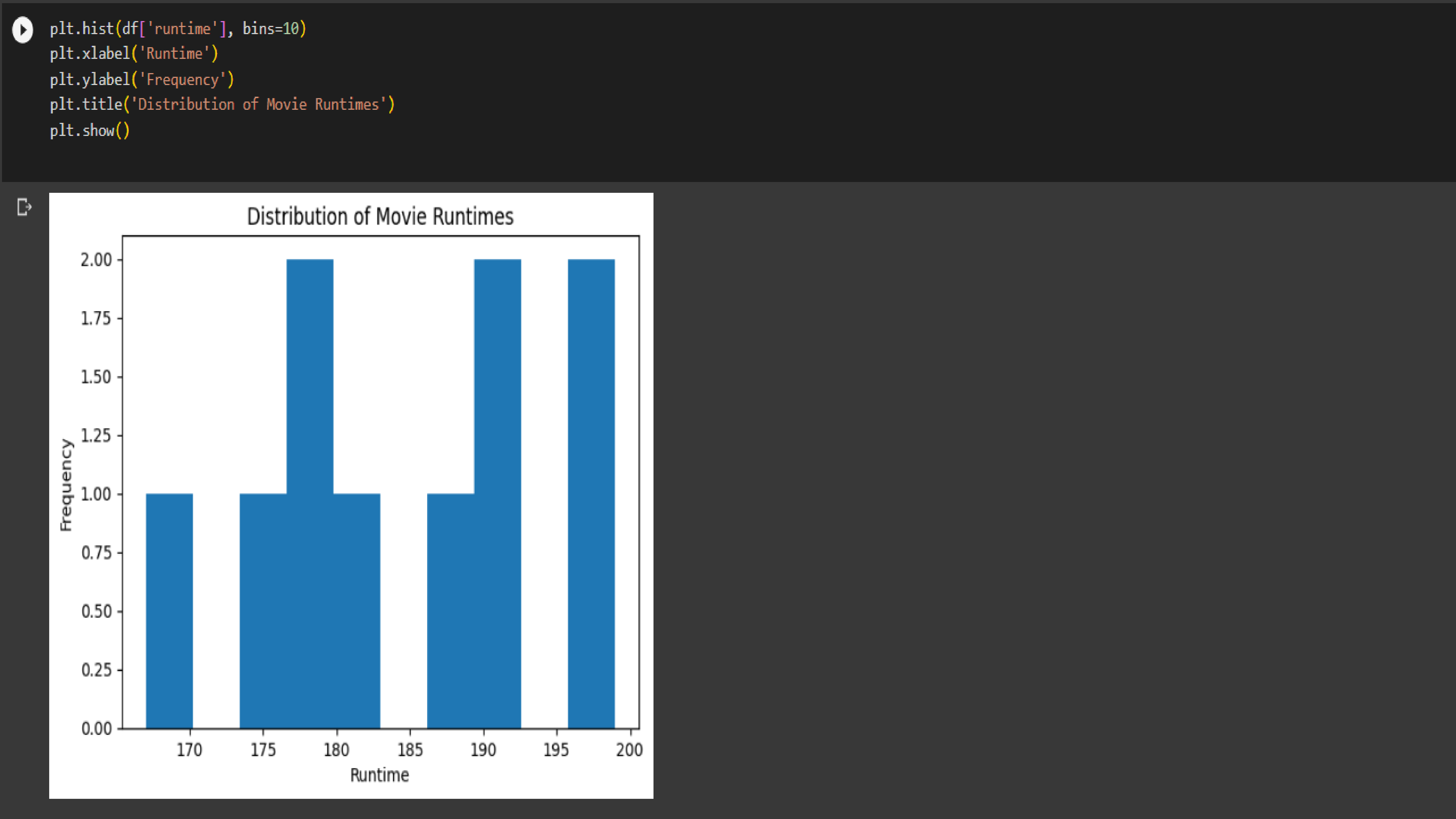
```
                    moviename   indusrty     budget  Domestic Gross  \
0     Avatar:The Way of water  Hollywood  460000000       684075767   
1                      Dangal  Bollywood    8536579        75975560   
2          Avengers : Endgame  Hollywood  400000000       858373000   
3  Baahubali2: The Conclusion  Tollywood   37000000        98067865   
4                      Fast X  Hollywood  340000000       145513495   
5                         RRR  Tollywood  720000000        56783625   
6    Avengers : Age of Ultron  Hollywood  365000000       459005868   
7             K.G.F Chapter 2  Tollywood   12195260        65745745   
8                     Pathaan  Bollywood   32927202        48368756   
9                    3 Idiots  Bollywood    6707393        13038765   

   Worldwide Gross  runtime   release date      Profit  
0       2320003887      198    Dec 9, 2022  1860003887  
1        243902280      178    Jan 13,2017   235365701  
2       2794731755      167   Apr 23 ,2019  2394731755  
3        143564874      190   Apr 28, 2017   106564874  
4        717245533      188   May 17 ,2023   377245533  
5        976545635      178    May 26,2018   256545635  
6       1395316979      181  Apr 22 , 2015  1030316979  
7        883475846      176   Jun 22, 2022   871280586  
8        864283654      199   Sep 13, 2022   831356452  
9        233294763      190   Dec 25, 2009   226587370  
```



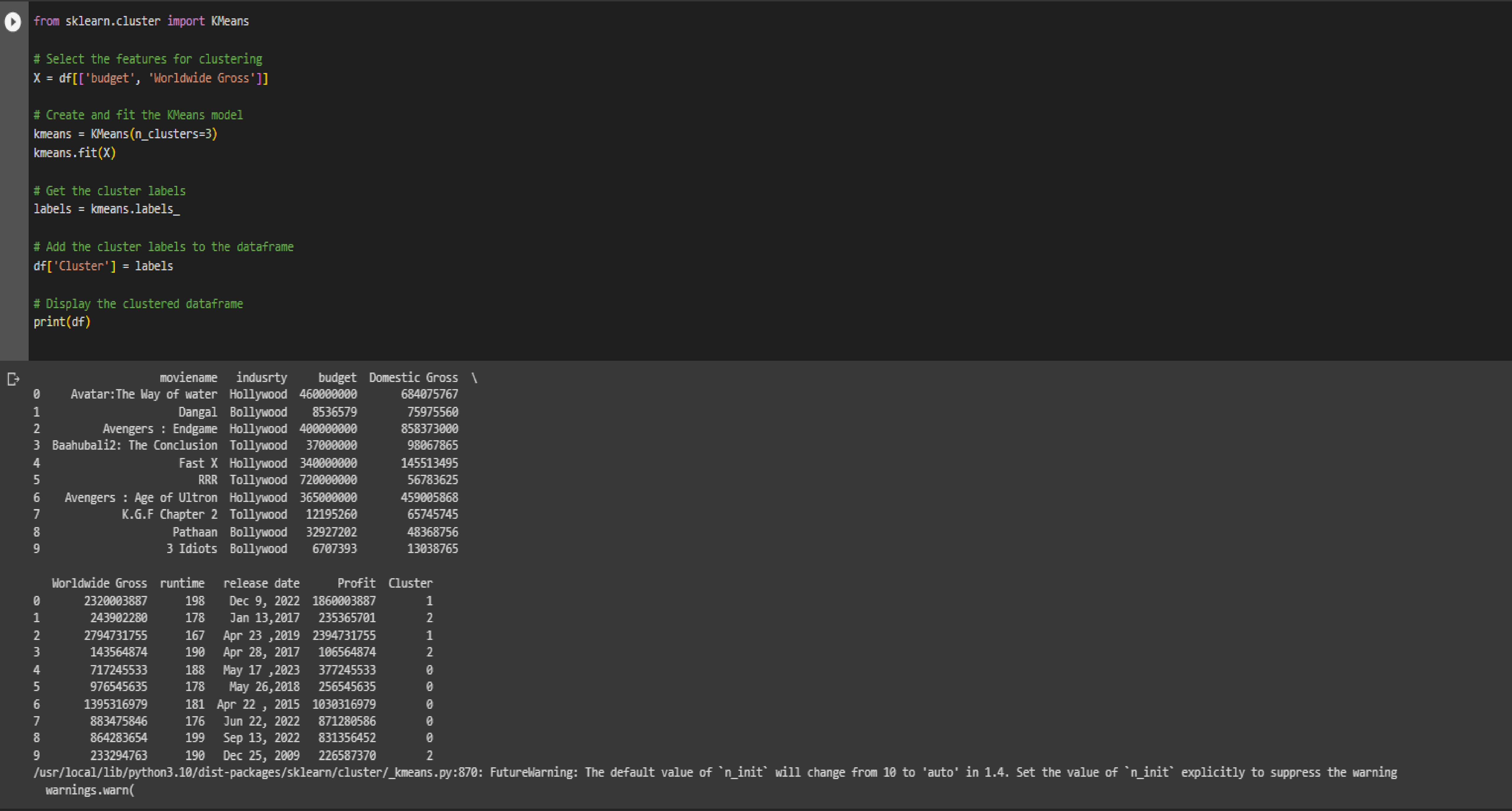
```
import matplotlib.pyplot as plt

plt.bar(df['indusrty'], df['budget'])
plt.xlabel('Indusrty')
plt.ylabel('Budget')
plt.title('Average Budget of Movies by Industry')
plt.show()
```



```
plt.scatter(df['budget'], df['Worldwide Gross'])
plt.xlabel('Budget')
plt.ylabel('Worldwide Gross')
plt.title('Budget vs Worldwide Gross')
plt.show()
```

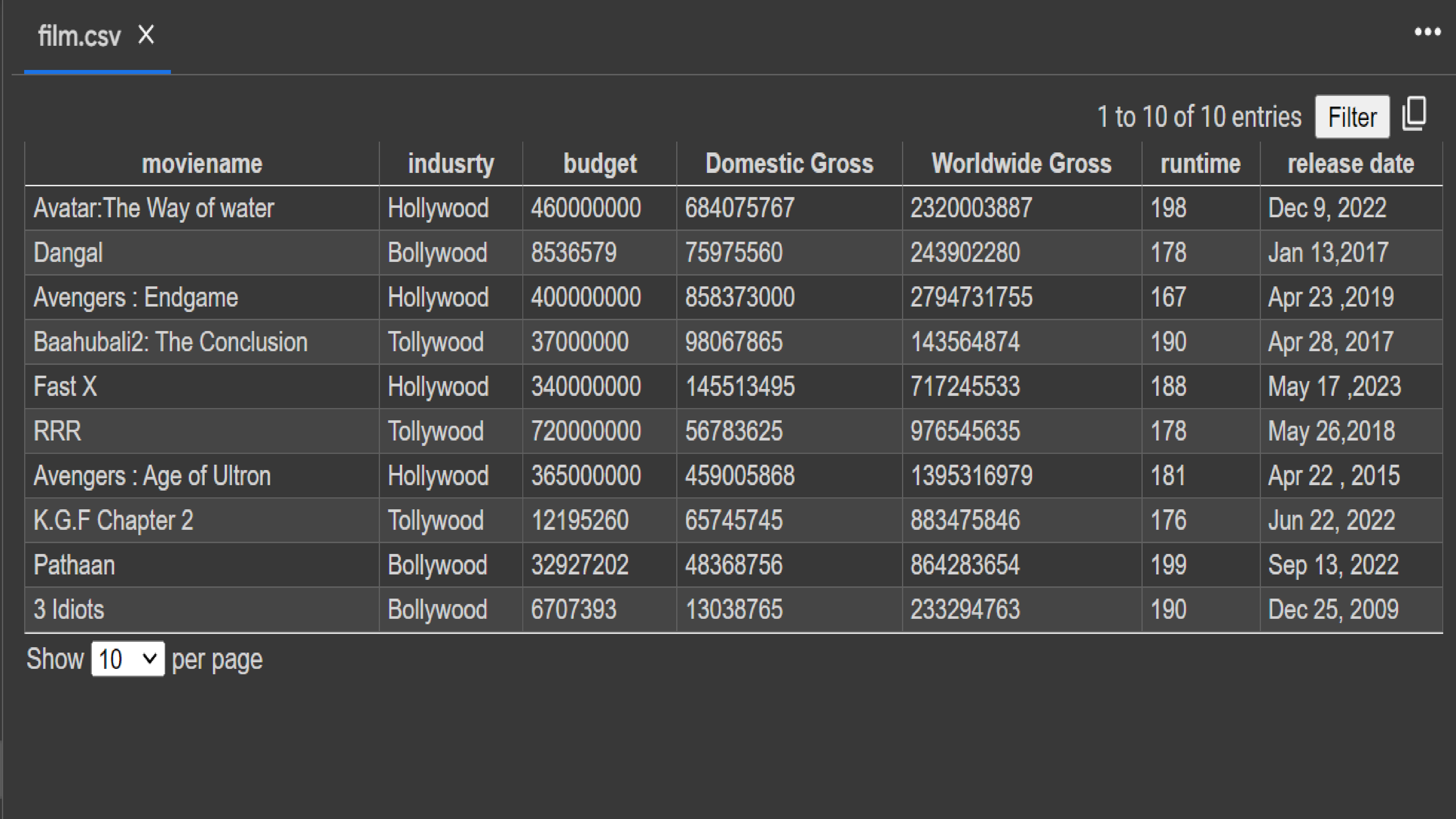


```
plt.hist(df['runtime'], bins=10)
plt.xlabel('Runtime')
plt.ylabel('Frequency')
plt.title('Distribution of Movie Runtimes')
plt.show()
```



```python
from sklearn.cluster import KMeans

# Select the features for clustering
X = df[['budget', 'Worldwide Gross']]

# Create and fit the KMeans model
kmeans = KMeans(n_clusters=3)
kmeans.fit(X)

# Get the cluster labels
labels = kmeans.labels_

# Add the cluster labels to the dataframe
df['Cluster'] = labels

# Display the clustered dataframe
print(df)
```

```
                    moviename   indusrty     budget  Domestic Gross  \
0     Avatar:The Way of water  Hollywood  460000000       684075767   
1                      Dangal  Bollywood    8536579        75975560   
2          Avengers : Endgame  Hollywood  400000000       858373000   
3  Baahubali2: The Conclusion  Tollywood   37000000        98067865   
4                      Fast X  Hollywood  340000000       145513495   
5                         RRR  Tollywood  720000000        56783625   
6    Avengers : Age of Ultron  Hollywood  365000000       459005868   
7             K.G.F Chapter 2  Tollywood   12195260        65745745   
8                     Pathaan  Bollywood   32927202        48368756   
9                    3 Idiots  Bollywood    6707393        13038765   

   Worldwide Gross  runtime   release date      Profit  Cluster  
0       2320003887      198    Dec 9, 2022  1860003887        1  
1        243902280      178    Jan 13,2017   235365701        2  
2       2794731755      167   Apr 23 ,2019  2394731755        1  
3        143564874      190   Apr 28, 2017   106564874        2  
4        717245533      188   May 17 ,2023   377245533        0  
5        976545635      178    May 26,2018   256545635        0  
6       1395316979      181  Apr 22 , 2015  1030316979        0  
7        883475846      176   Jun 22, 2022   871280586        0  
8        864283654      199   Sep 13, 2022   831356452        0  
9        233294763      190   Dec 25, 2009   226587370        2  
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```



| moviename | indusrty | budget | Domestic Gross | Worldwide Gross | runtime | release date |
|---|---|---|---|---|---|---|
| Avatar:The Way of water | Hollywood | 460000000 | 684075767 | 2320003887 | 198 | Dec 9, 2022 |
| Dangal | Bollywood | 8536579 | 75975560 | 243902280 | 178 | Jan 13,2017 |
| Avengers : Endgame | Hollywood | 400000000 | 858373000 | 2794731755 | 167 | Apr 23 ,2019 |
| Baahubali2: The Conclusion | Tollywood | 37000000 | 98067865 | 143564874 | 190 | Apr 28, 2017 |
| Fast X | Hollywood | 340000000 | 145513495 | 717245533 | 188 | May 17 ,2023 |
| RRR | Tollywood | 720000000 | 56783625 | 976545635 | 178 | May 26,2018 |
| Avengers : Age of Ultron | Hollywood | 365000000 | 459005868 | 1395316979 | 181 | Apr 22 , 2015 |
| K.G.F Chapter 2 | Tollywood | 12195260 | 65745745 | 883475846 | 176 | Jun 22, 2022 |
| Pathaan | Bollywood | 32927202 | 48368756 | 864283654 | 199 | Sep 13, 2022 |
| 3 Idiots | Bollywood | 6707393 | 13038765 | 233294763 | 190 | Dec 25, 2009 |