Data Analysis Report: Movie Dataset

1. Objective

The primary goal of this analysis is to explore the relationship between **budget** and **gross earnings** in a movie dataset. The analysis includes:

- Data cleaning
- Missing value handling
- Data visualization
- Correlation analysis

2. Libraries Used

- `pandas` – for data manipulation
- `numpy` – for numerical operations
- `matplotlib.pyplot` & `seaborn` – for visualization
- `matplotlib` – for plot configuration

3. Data Loading & Initial View

```
df = pd.read_csv(r"C:\Users\tamil\Desktop\Tamil\DATA ANALYST\Python\Movies\movies.csv")
df.head()
```

4. Handling Missing Data

```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, pct_missing))
```

This step identifies the percentage of missing values in each column.

Missing Data Handling

df['rating'] = df['rating'].fillna(0).astype(object)

df['company'] = df['company'].fillna(0).astype(object)

Instead of dropping, missing values are replaced with `0`, and data types are cast to `object`.

5. Sorting and Duplicate Removal

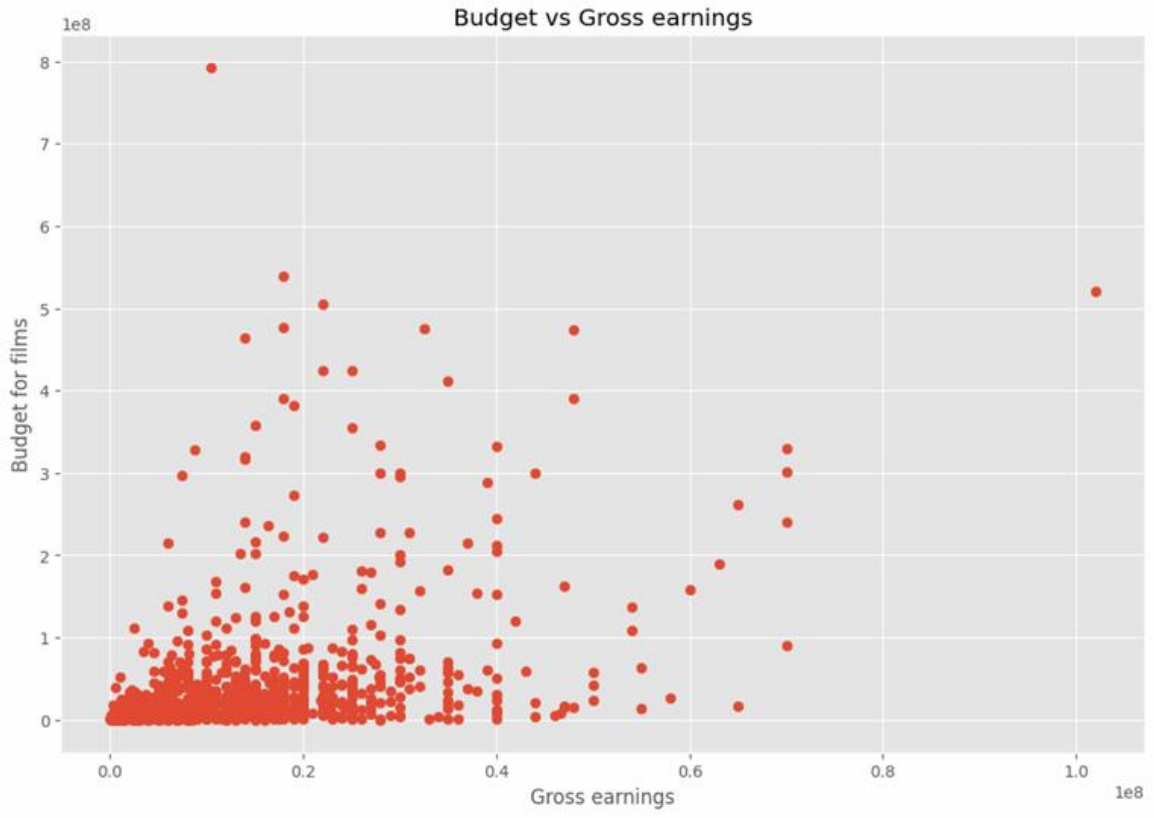
df.sort_values(by=['gross'], ascending=False)

df.drop_duplicates()

The dataset is sorted by the `gross` column to prioritize high-earning movies. Duplicate rows are dropped.

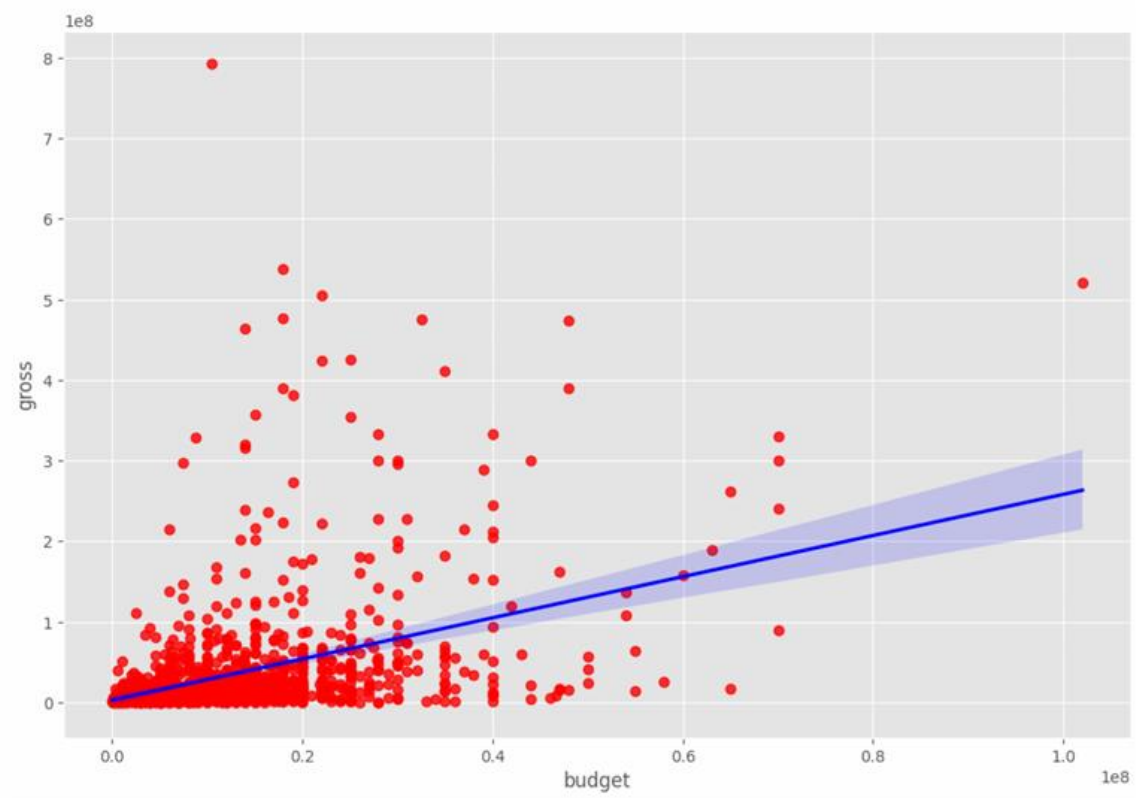
6. Data Visualization

Scatter Plot: Budget vs Gross



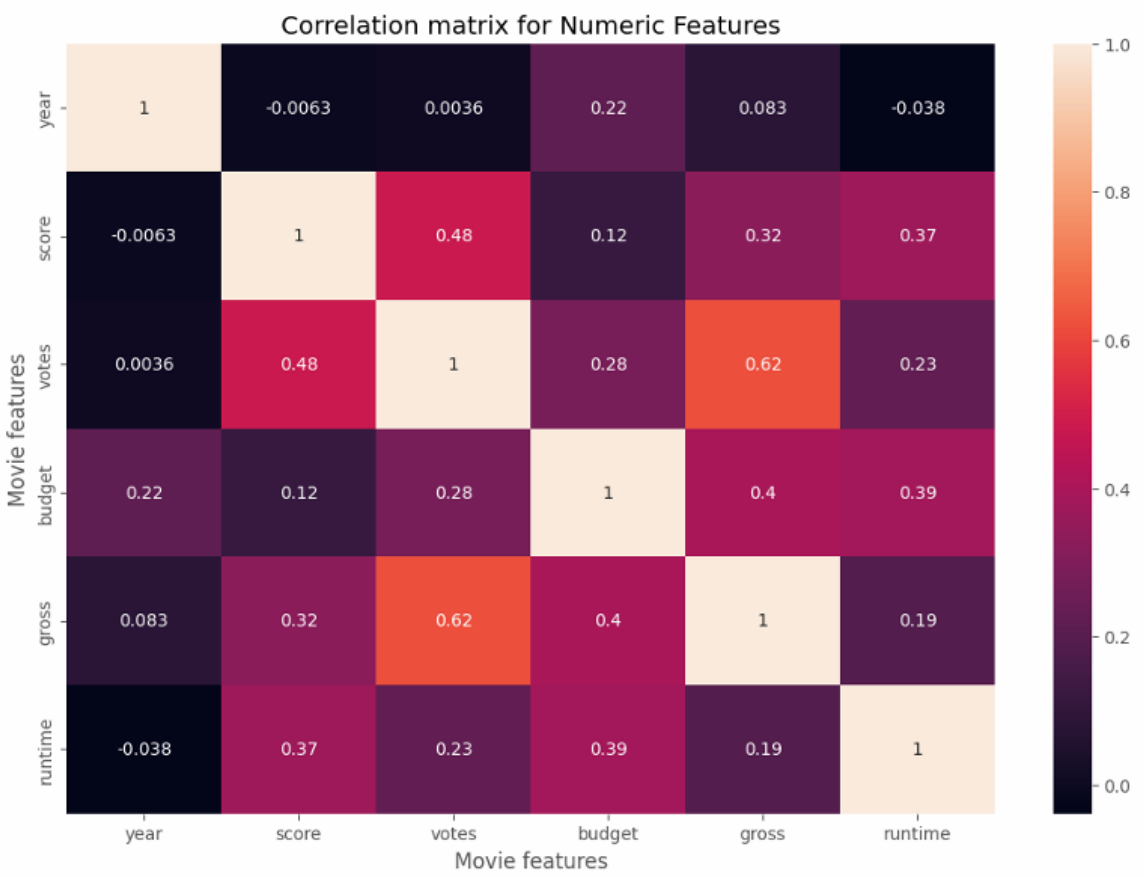
A scatter plot shows the relationship between movie budget and gross revenue.

Regression Plot using Seaborn



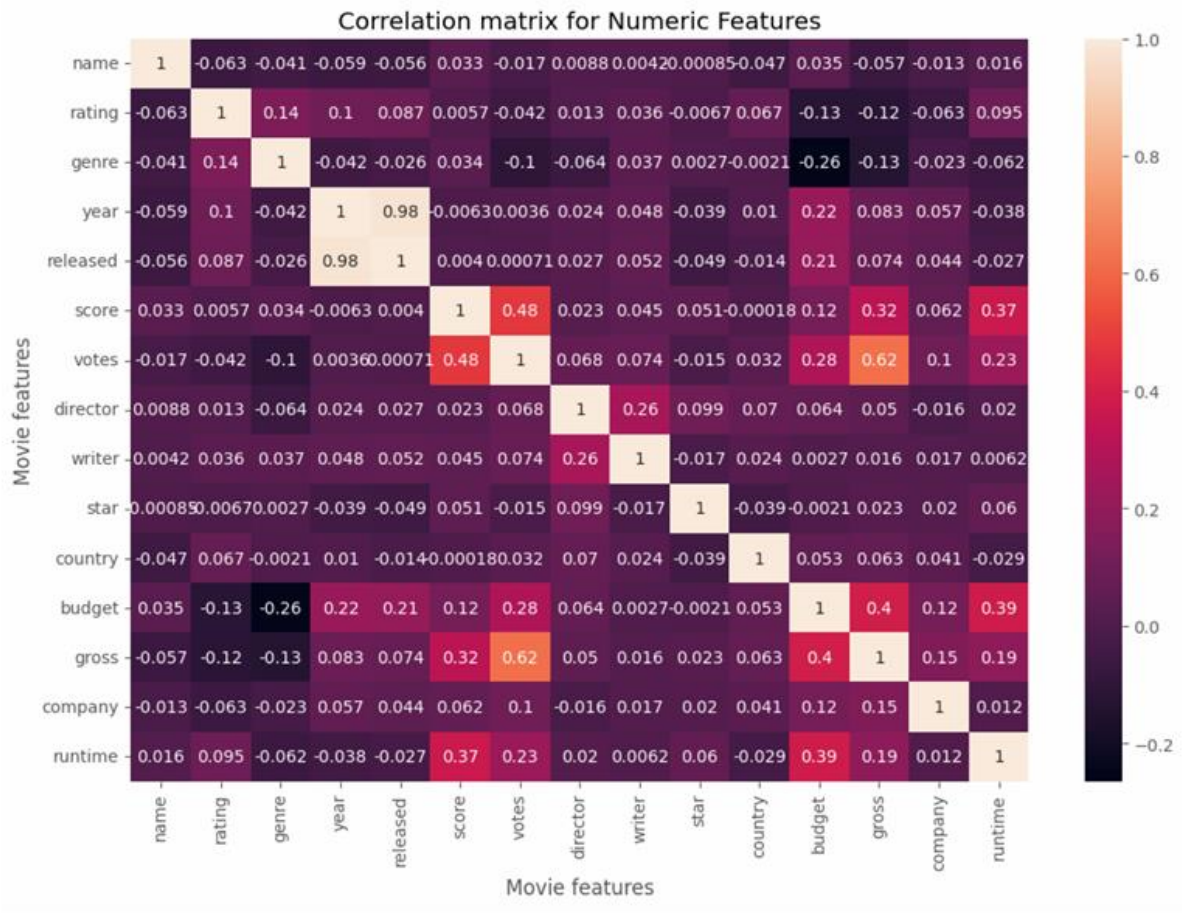
7. Correlation Analysis

Numerical Correlation (Pearson)



This visualizes how numeric features like budget and gross are correlated.

Categorical Conversion & Full Correlation



This encodes categorical columns numerically so they can be included in correlation analysis.

High Correlation Pair Extraction

corr_pairs = correlation_mat.unstack()

sorted_pairs = corr_pairs.sort_values()

high_corr = sorted_pairs[(sorted_pairs) > 0.5]

8. Key Insights

- **Budget and Gross** are **strongly correlated**, meaning higher budget tends to result in higher gross earnings.
- Other features might also show moderate-to-strong correlations after encoding.