Dataset Overview

This dataset comprises information on 10,867 movies released between 1960 and 2015. Each movie is uniquely identified by `id` and `imdb_id` and includes details such as budget, revenue (original and inflation-adjusted), cast, director, genre, release date, user ratings (`vote_count`, `vote_average`), and other metadata. The data's primary focus appears to be on financial and critical success metrics, complemented by descriptive information.

Quality Assessment

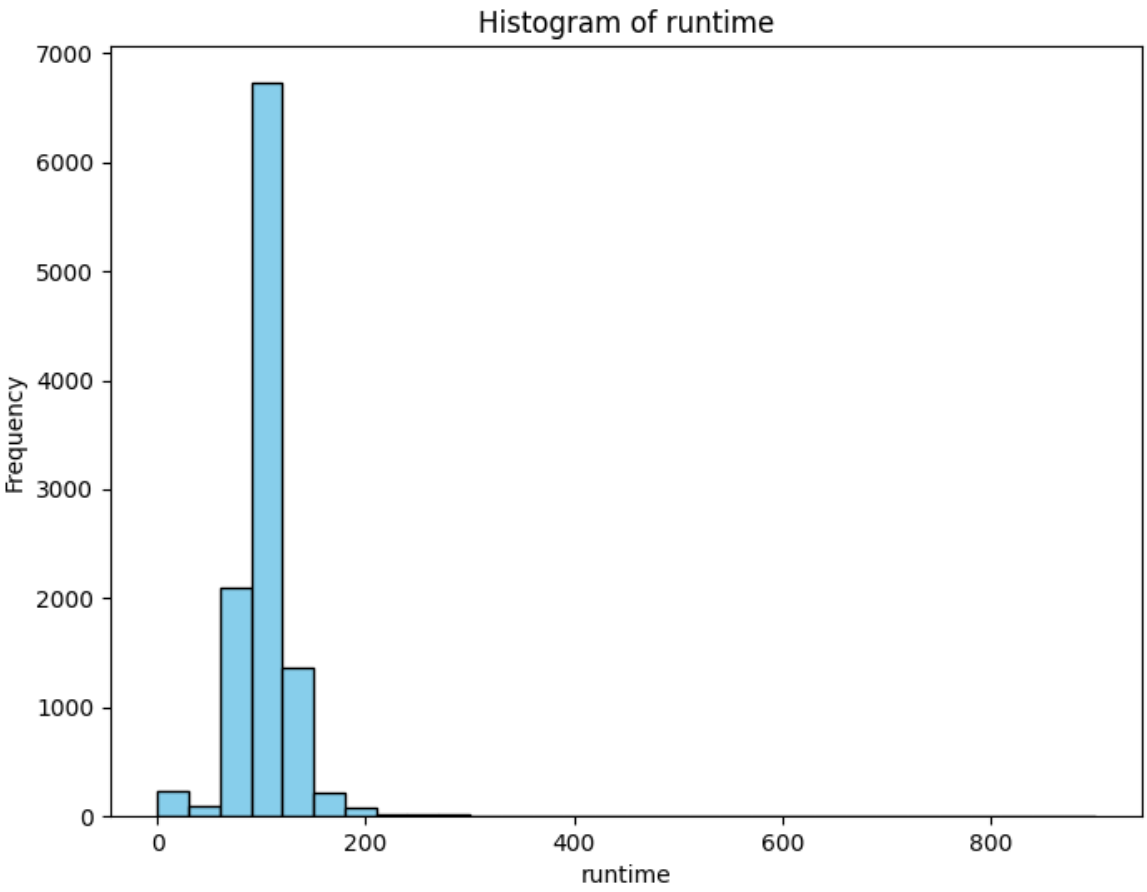
The dataset demonstrates a relatively high level of completeness. The `id` column lacks missing values, serving as a reliable primary key. However, several columns contain missing values: `imdb_id` (10 missing), `cast` (76 missing), `homepage` (7931 missing), `director` (44 missing), `tagline` (2824 missing), `keywords` (1493 missing), and `genres` (23 missing). The substantial number of missing values in `homepage` is expected, as not all movies possess official websites. Missing values in other columns warrant further investigation to determine the cause and consider imputation or removal strategies. The presence of 0 values in `budget` and `revenue` likely indicates unavailable data rather than a true zero value, requiring careful consideration during analysis.

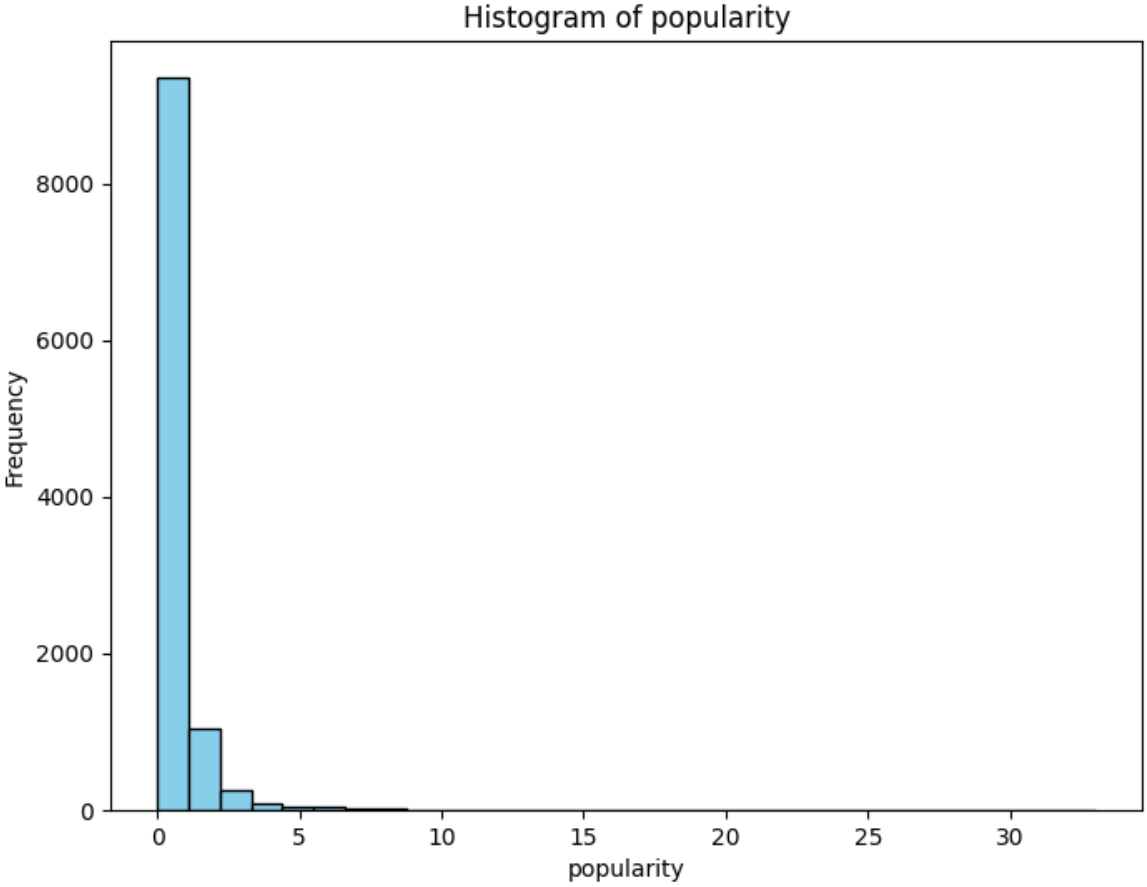
Descriptive Statistics

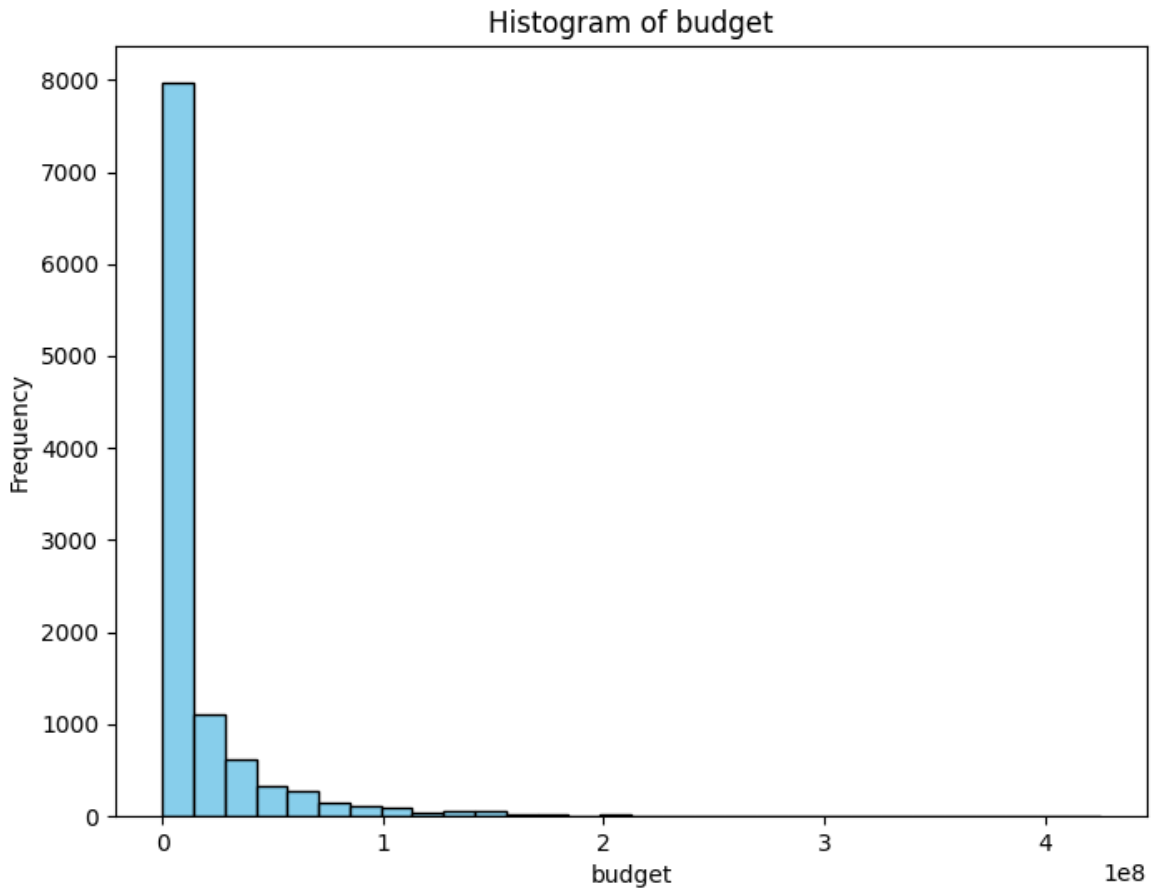
The following table summarizes key numerical variables:

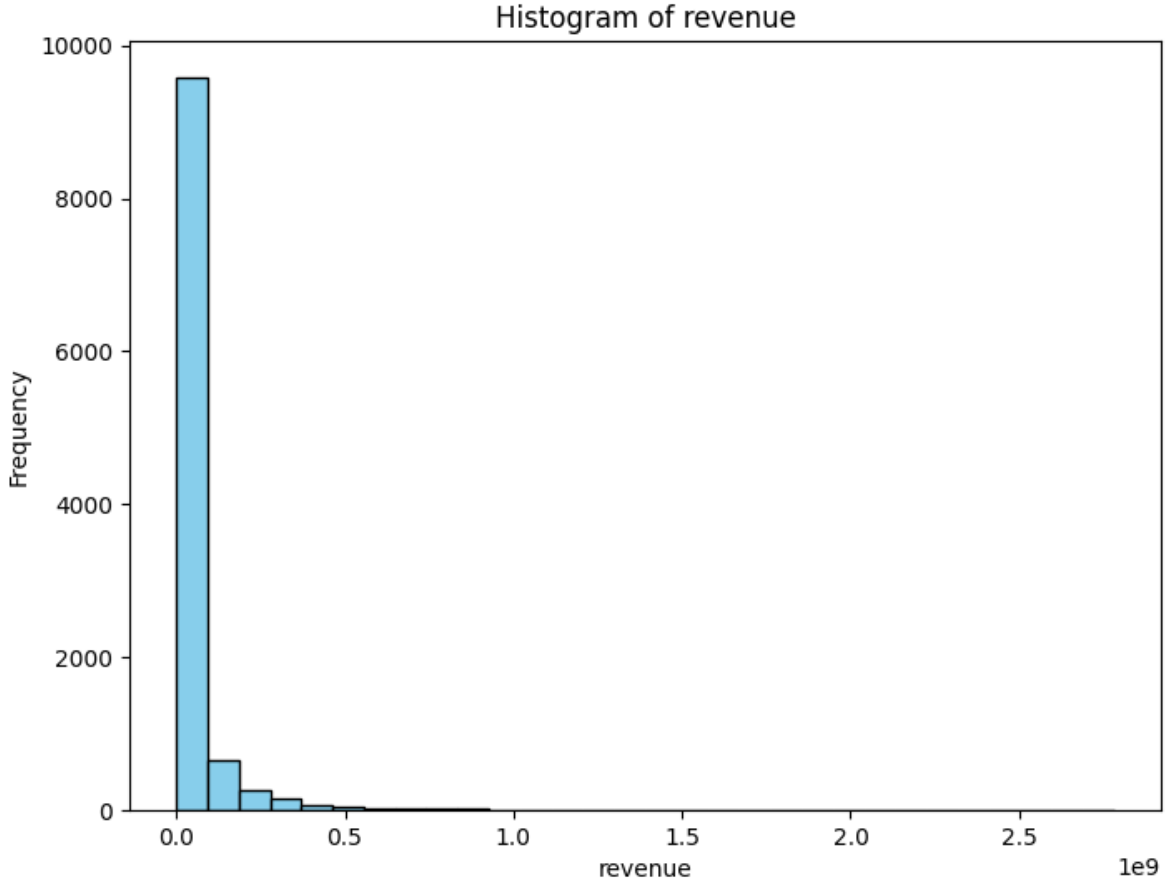
Variable	Mean	Median	Std Dev	Min	Max	25th Percentile	75th Percentile
budget	1.46e+07	0	3.09e+07	0	4.25e+08	0	1.50e+07
budget_adj	1.75e+07	0	3.43e+07	0	4.25e+08	0	2.09e+07
revenue	3.98e+07	0	1.17e+08	0	2.78e+09	0	2.40e+07
revenue_adj	5.14e+07	0	1.45e+08	0	2.83e+09	0	3.37e+07
popularity	0.65	0.38	1.00	0.000065	32.99	0.21	0.71
runtime	102.07	99	31.38	0	900	90	111
vote_average	5.97	6	0.94	1.5	9.2	5.4	6.6
vote_count	217.37	38	575.60	10	9767	17	145.5

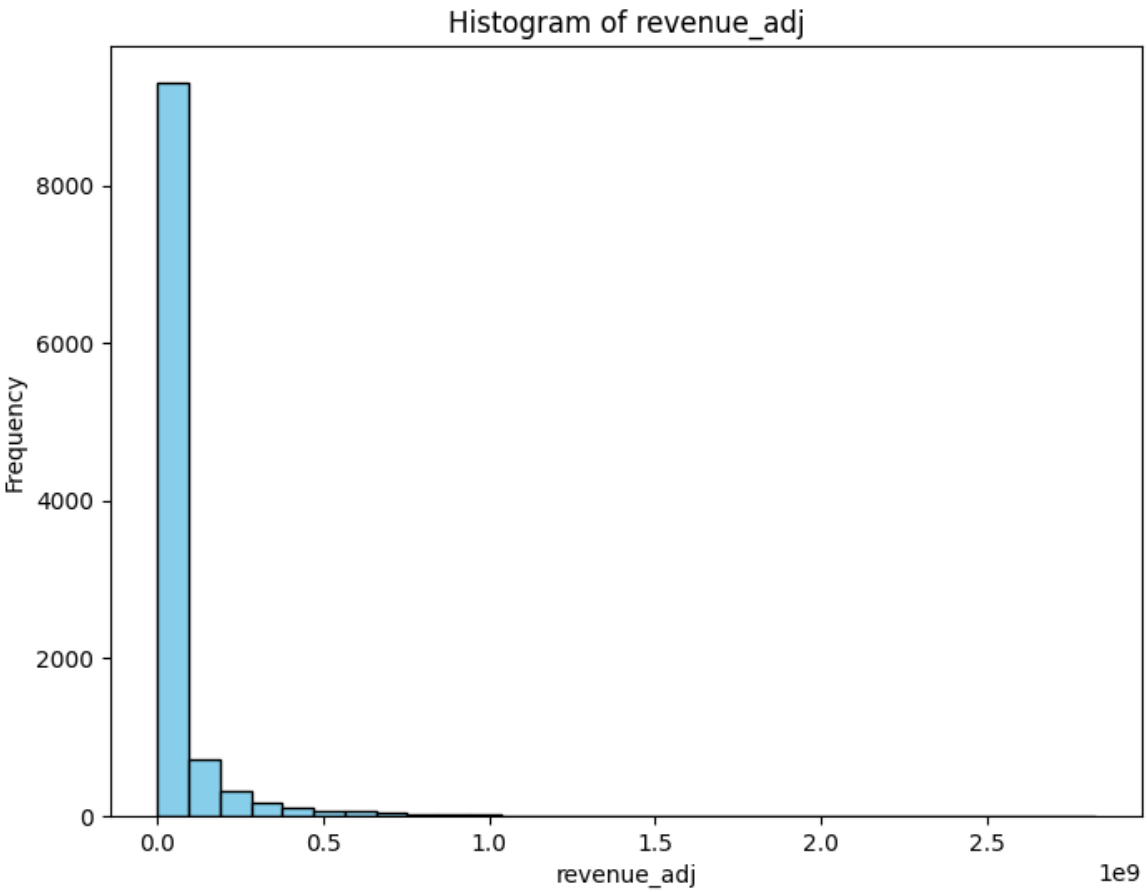
Histograms for all these variables are provided below:

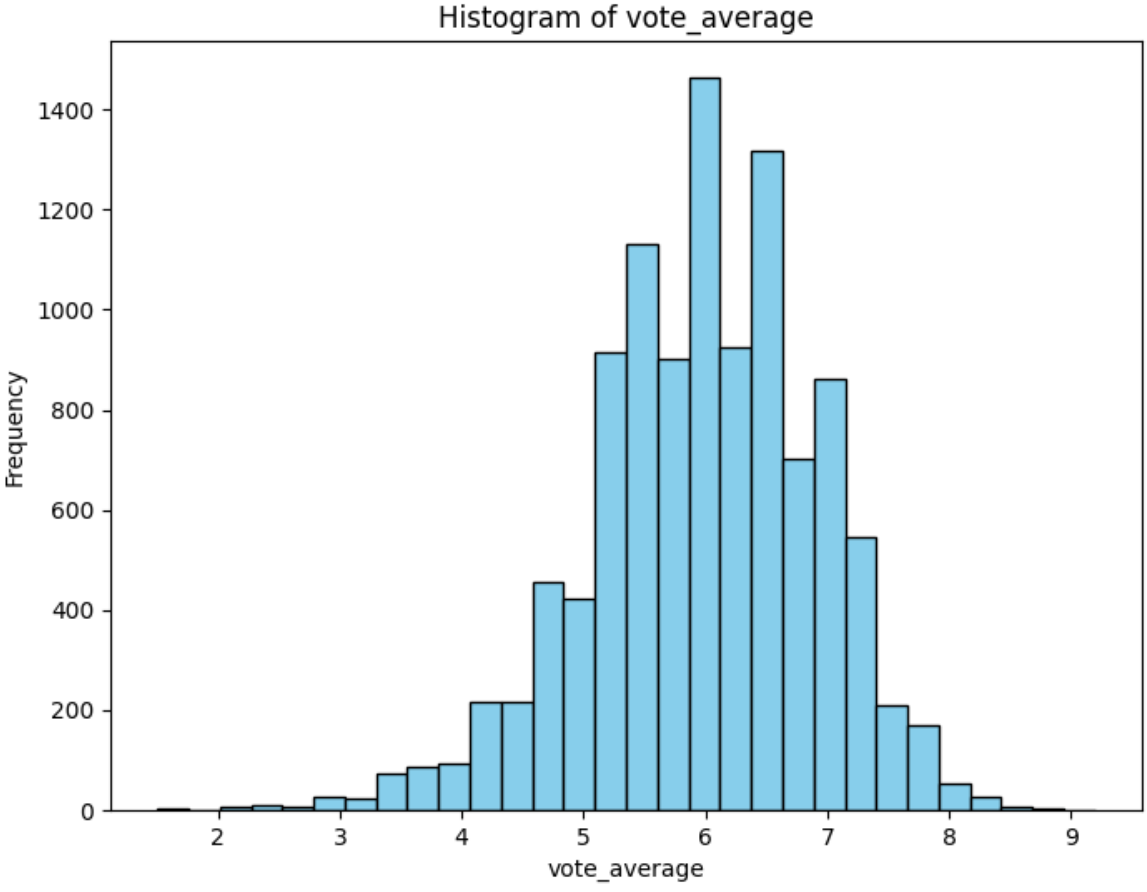

Histogram of runtime
Frequency
runtime
0
1000
2000
3000
4000
5000
6000
7000
0
200
400
600
800

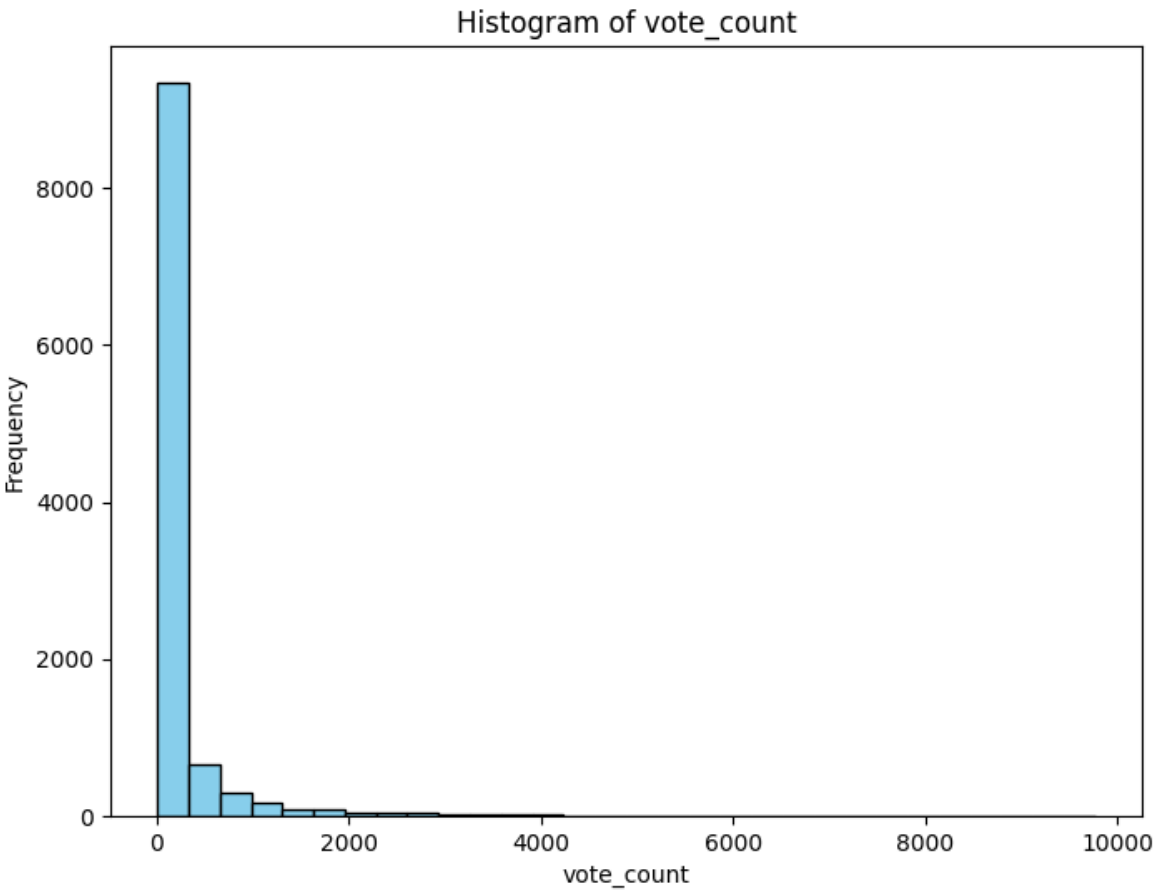

Histogram of popularity
Frequency
popularity
0
2000
4000
6000
8000
0
5
10
15
20
25
30

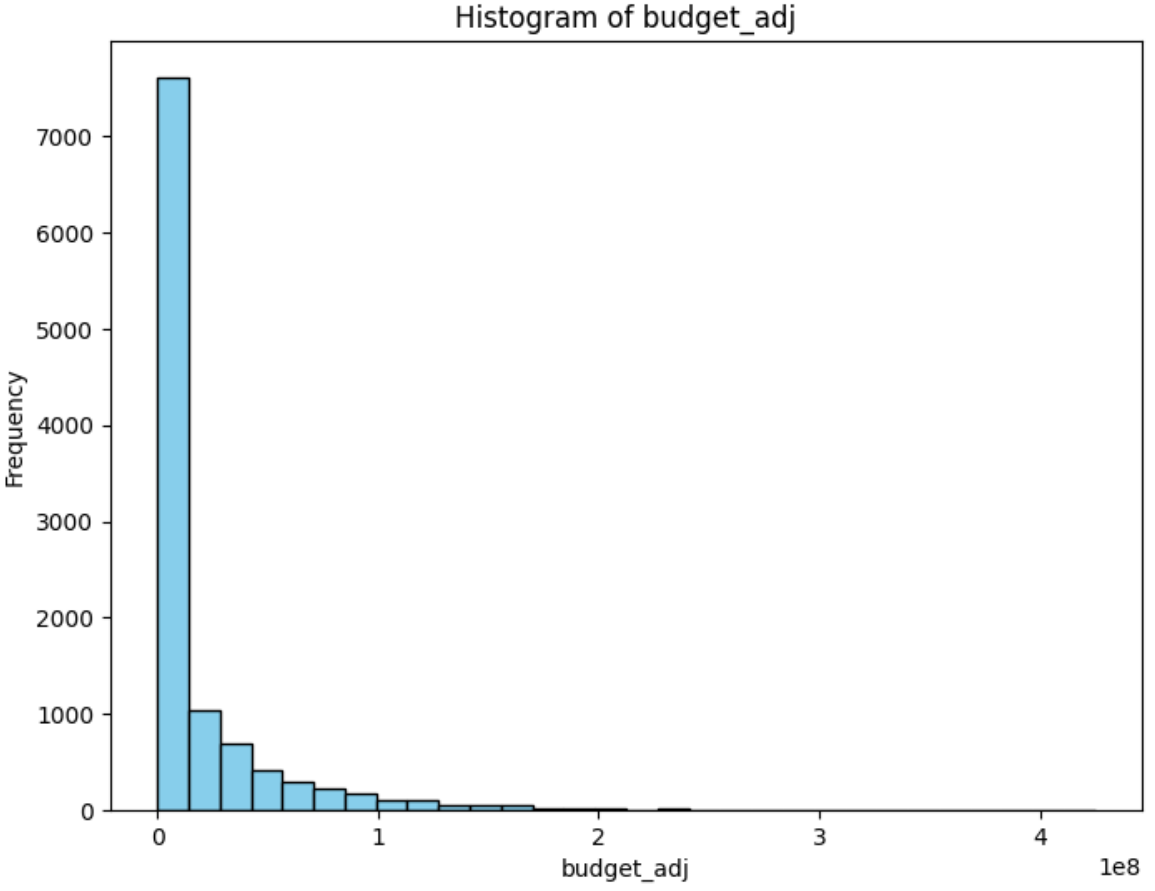

Histogram of budget
Frequency
budget
1e8
0
1000
2000
3000
4000
5000
6000
7000
8000
0
1
2
3
4


Histogram of revenue
Frequency
revenue
1e9
0
2000
4000
6000
8000
10000
0.0
0.5
1.0
1.5
2.0
2.5


Histogram of revenue_adj
Frequency
8000
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
2.5
revenue_adj
1e9


Histogram of vote_average
Frequency
1400
1200
1000
800
600
400
200
0
2
3
4
5
6
7
8
9
vote_average


Histogram of vote_count
Frequency
vote_count
0
2000
4000
6000
8000
10000

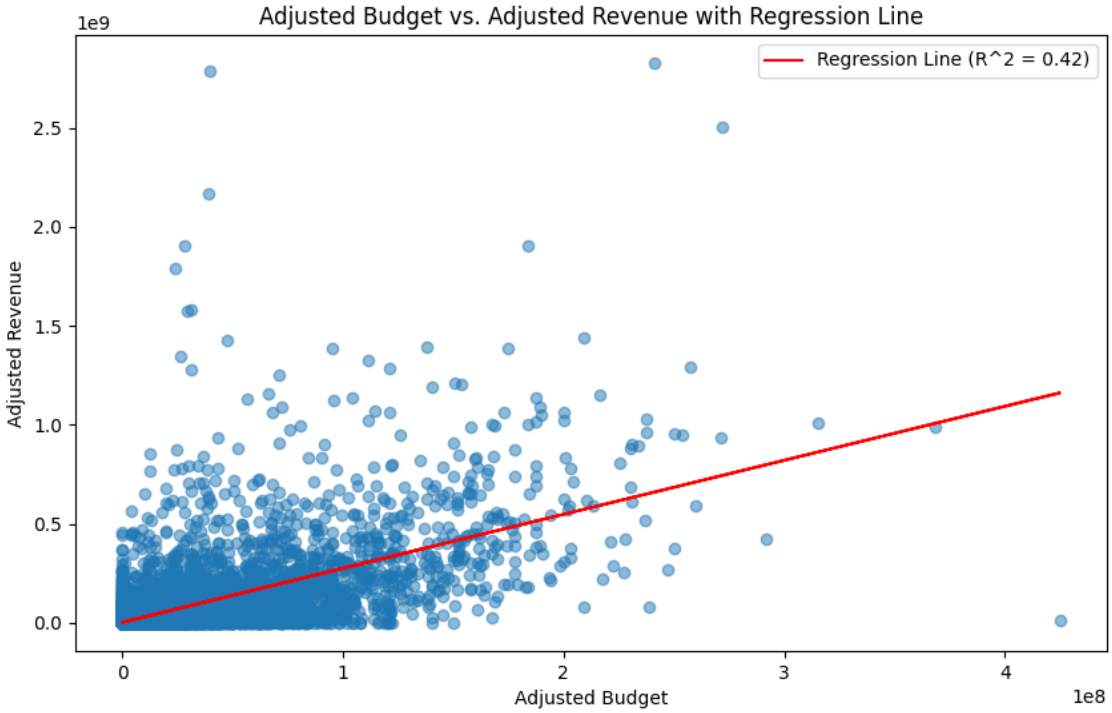

Histogram of budget_adj
Frequency
budget_adj
1e8
0
1000
2000
3000
4000
5000
6000
7000
0
1
2
3
4

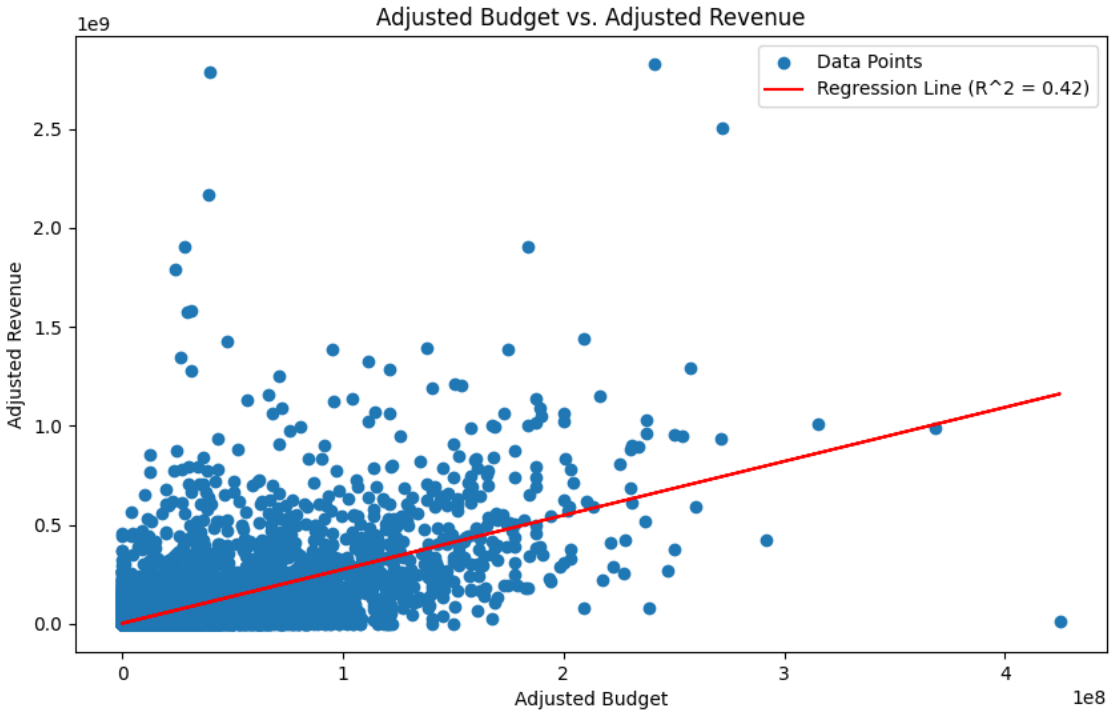
The most frequent genre is Drama (4761 movies), followed by Comedy (3793). Woody Allen directed the most movies (45), followed by Clint Eastwood (34).

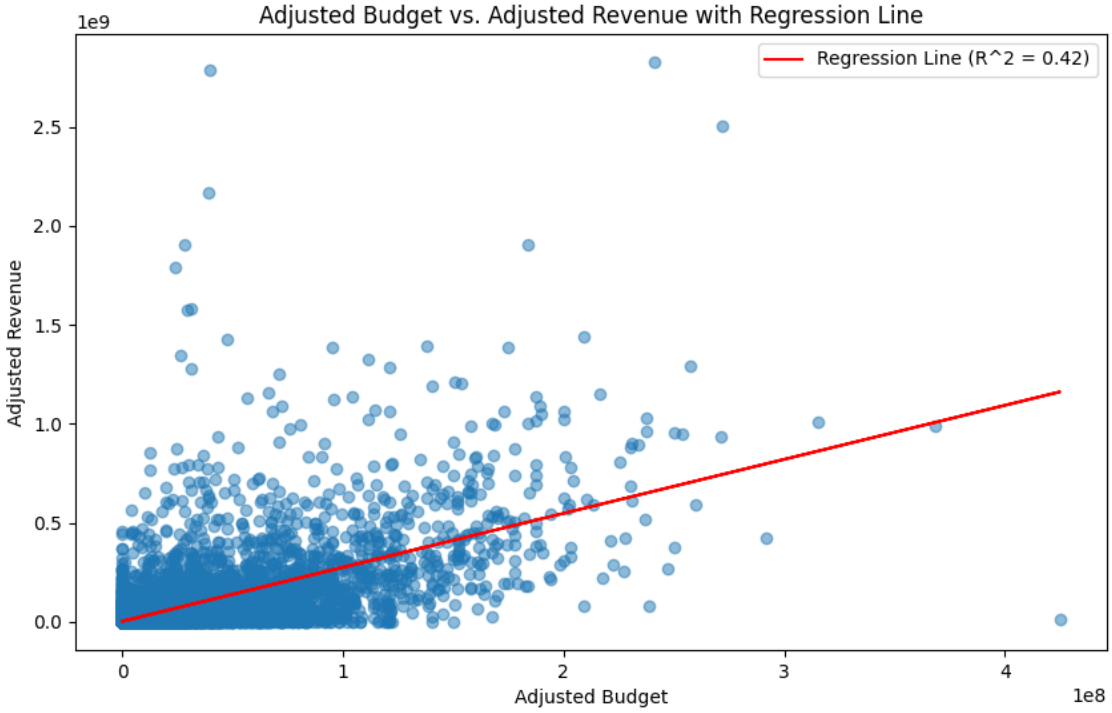
EDA

Budget vs. Revenue

The scatter plot with regression line (see below) shows a positive correlation between adjusted budget and adjusted revenue, as expected. However, there's significant scatter, indicating other factors influence revenue beyond budget. Note that the attempt to generate this plot initially encountered errors; the provided images represent different attempts to generate the plot.

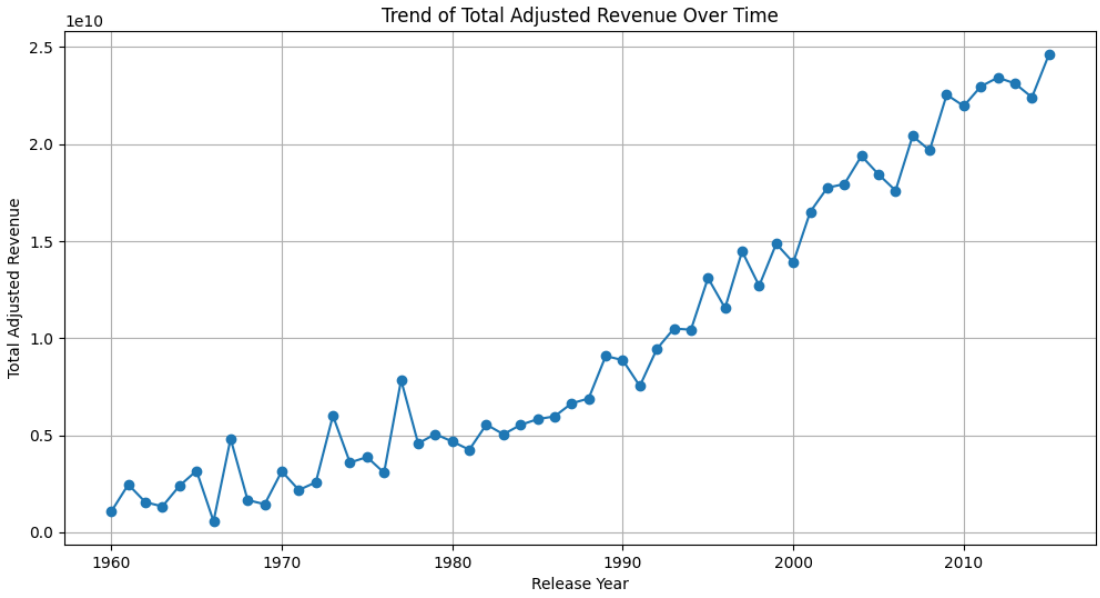


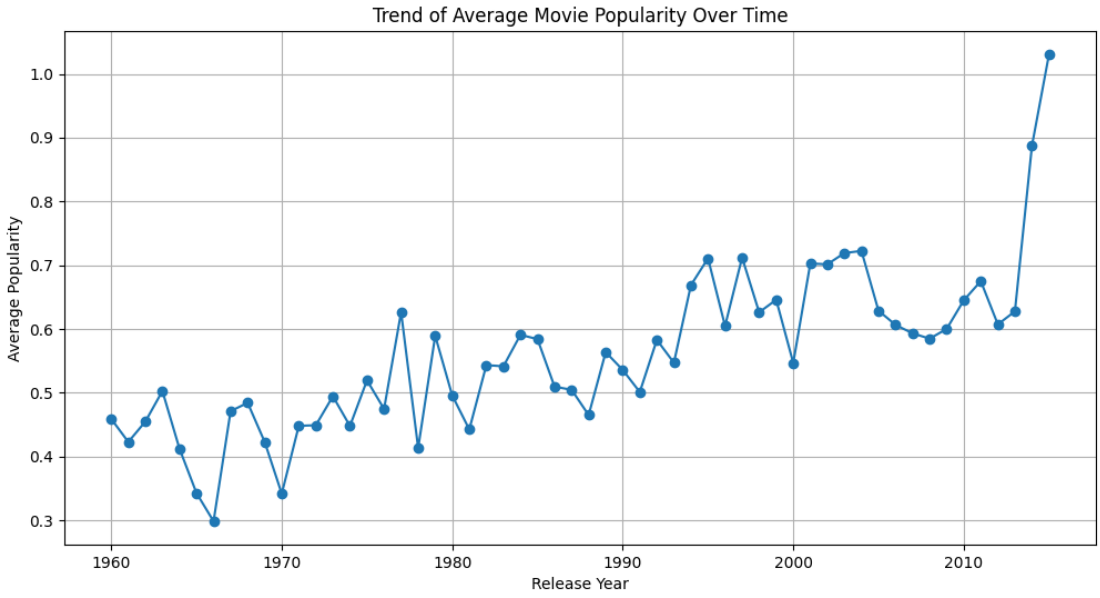




Revenue and Popularity Over Time

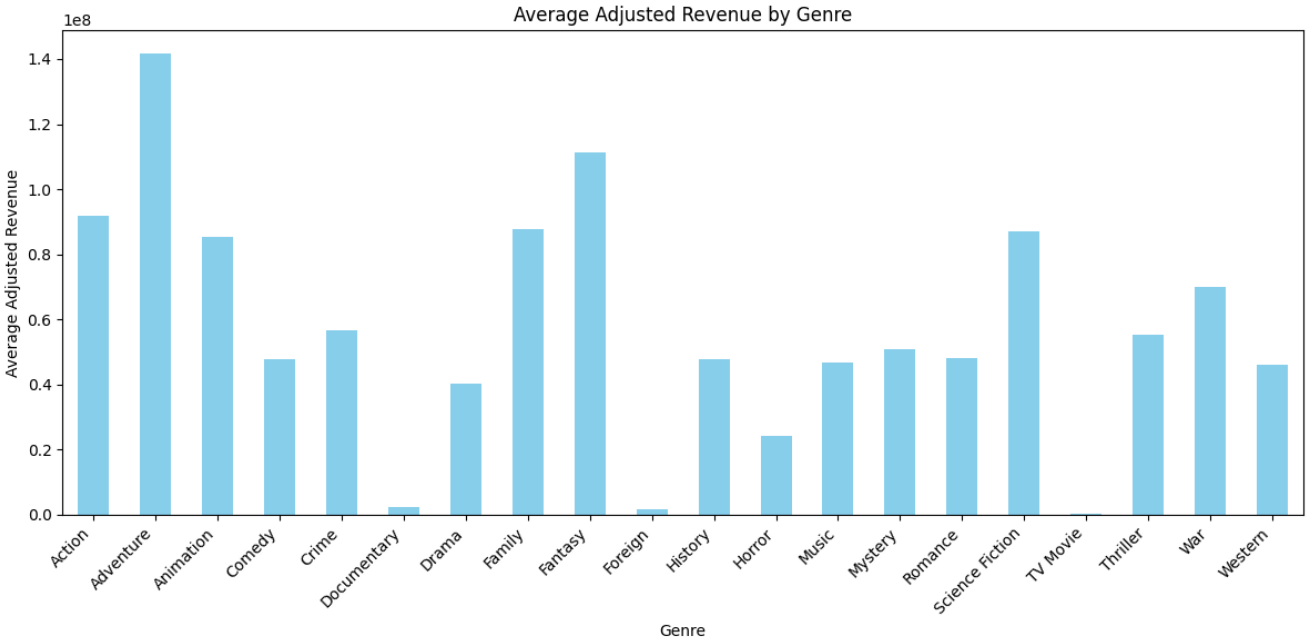
The line charts below illustrate the trends of total adjusted revenue and average popularity over the years. Both show a general upward trend, suggesting growth in both revenue and popularity of movies over time.

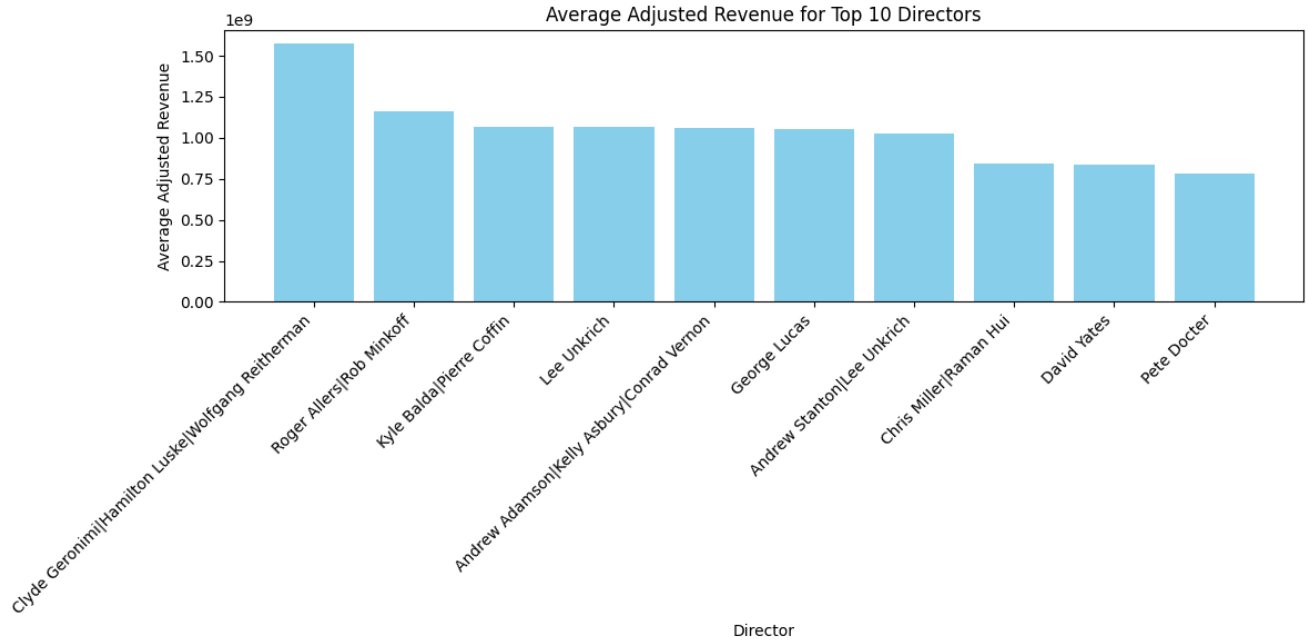




Genre and Director Revenue

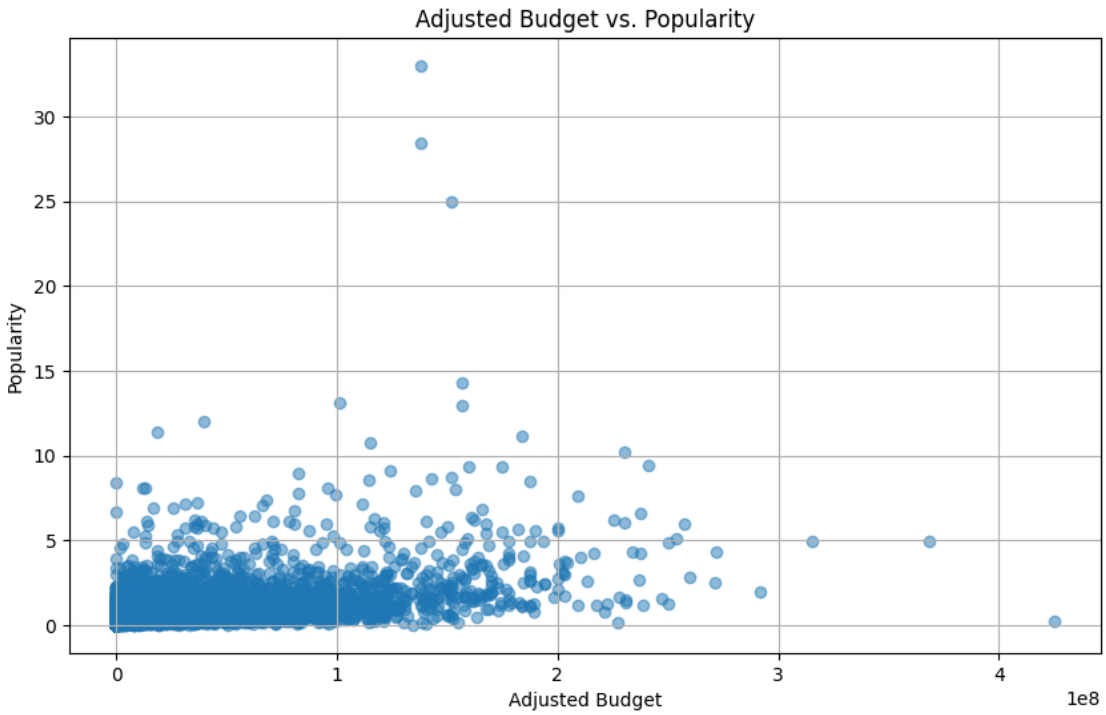
The bar charts below show the average adjusted revenue per genre and for the top 10 directors. Adventure and Science Fiction genres have the highest average revenue, while Documentary and TV Movie have the lowest. Clyde Geronimi, Hamilton Luske, and Wolfgang Reitherman have the highest average revenue among the top 10 directors.





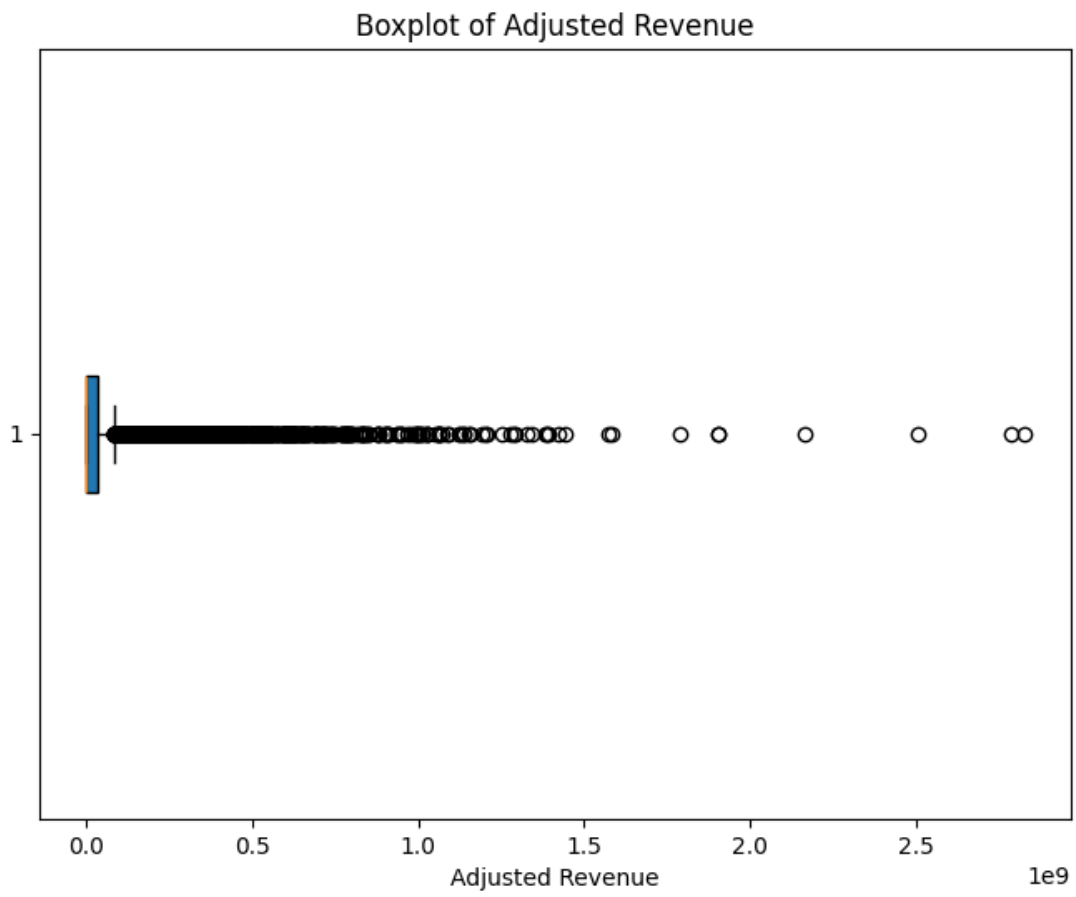
Budget vs. Popularity

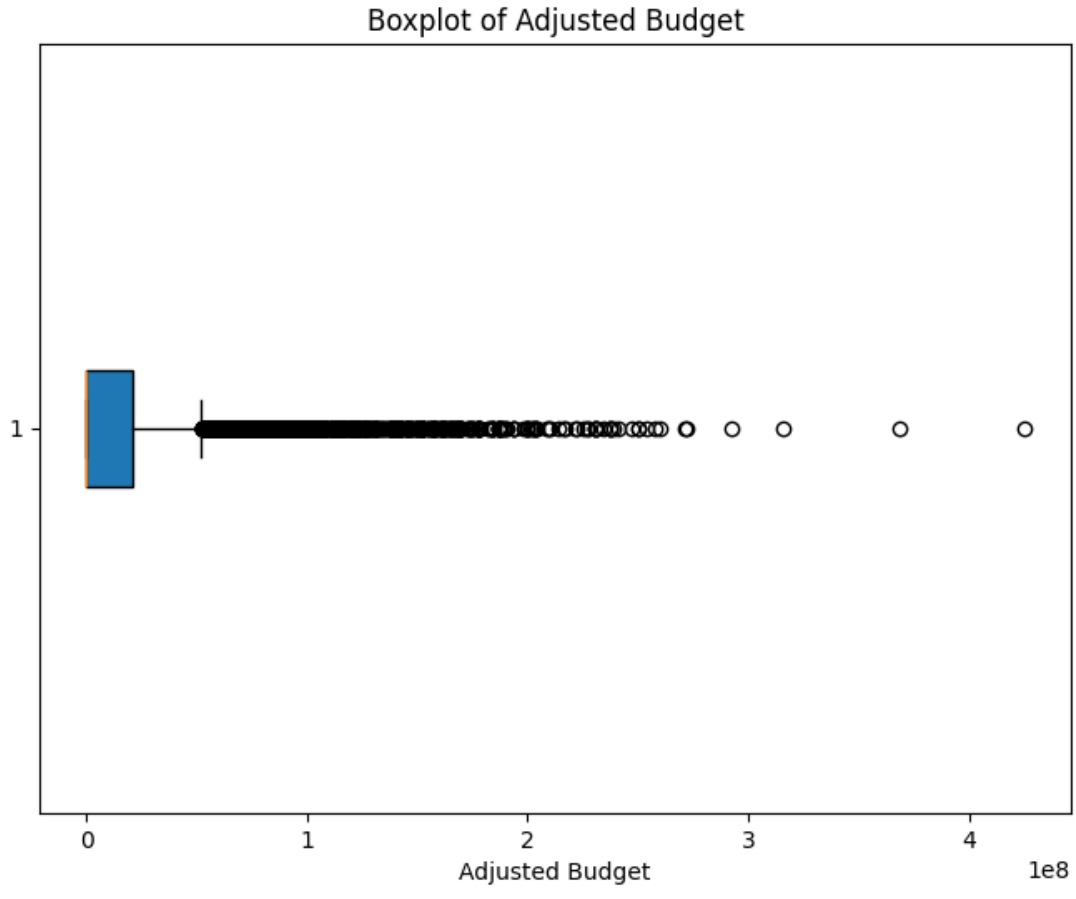
The scatter plot below shows a positive correlation between adjusted budget and popularity, suggesting that higher-budget movies tend to be more popular. However, the relationship is not perfectly linear, and many high-budget movies have relatively low popularity scores.



Boxplots of Budget and Revenue

The boxplots below show the distribution of adjusted budget and adjusted revenue, highlighting the presence of many outliers (movies with exceptionally high budgets and revenues).


Boxplot of Adjusted Revenue
1
0.0
0.5
1.0
1.5
2.0
2.5
Adjusted Revenue
1e9


Boxplot of Adjusted Budget
1
0
1
2
3
4
Adjusted Budget
1e8

Key Findings

- A strong positive correlation exists between adjusted budget and adjusted revenue, although other factors significantly influence revenue.
- Both total adjusted revenue and average movie popularity have generally increased over time.
- Adventure and Science Fiction genres tend to generate higher average revenues.
- Certain directors consistently deliver high-revenue movies.
- Significant outliers exist in both adjusted budget and adjusted revenue, warranting further investigation.

Summary

This analysis reveals a positive correlation between movie budget and revenue, along with a general upward trend in both revenue and popularity over time. Genre and director significantly impact revenue, with certain genres and directors consistently achieving higher financial success. The presence of outliers suggests the existence of exceptionally successful or unsuccessful movies that warrant further investigation. Further analysis could explore the impact of other variables (e.g., cast, keywords, release date) on movie success.