EDA Report

Business Objectives

GOALS:

- Find insights into popular movies and how things like budget, revenue, genre, director, etc. affect each other

Basic Characteristics

From two files; One Crew one Movies

Data characteristics:

- 24 variables across two files
- Movies file: 20 columns
  - Financials: Budget, Revenue
  - Descriptions: Title, Overview, Tagline
  - Release Information: Release Date, Status
  - Popularity Metrics: Popularity score, Vote Average, Vote Count
- Credits file: 4 columns
  - Title, movie ID, cast (JSON), crew (JSON)

4803 Rows across both

Movies Data Dictionary

Variable Name	Type	Description
id	Integer	Unique ID for each movie
imdb_id	String	IMDB ID for the movie
title	String	Title of the movie
original_title	String	Original language title of the movie
original_language	String	ISO 639-1 code for the original language
overview	String	Brief description or summary
tagline	String	Short marketing tagline
genres	String (JSON)	List of genres associated with the movie
release_date	Date	Date when the movie was first released
runtime	Float	Duration of the movie in minutes

Movies Data Dictionary

Variable Name	Type	Description
budget	Float	Budget of the movie (in USD)
revenue	Float	Revenue generated by the movie (in USD)
popularity	Float	Popularity score assigned by TMDb
vote_average	Float	Average user rating
vote_count	Integer	Number of user ratings submitted
status	String	Release status (e.g., Released, Post Production)
homepage	String	Official website URL
production_companies	String (JSON)	List of production companies involved
production_countries	String (JSON)	List of countries where production occurred
spoken_languages	String (JSON)	Languages spoken in the movie

Crew Data Dictionary

Variable Name	Type	Description
id	Integer	Unique ID for each movie (matches movies_metadata.id)
title	String	Title of the movie
cast	String (JSON)	List of cast members and their roles
crew	String (JSON)	List of crew members and their jobs

Variable Classifications (Crew)

Crew:

- movie_id - Nominal Category
- title - Nominal Category
- cast - JSON list
- crew - JSON list

Variable Classifications (Movies)

budget - Measure (ratio)
genres - JSON List
homepage - Nominal Category
id - Nominal Category
keywords - JSON List
original_language - Nominal Category
original_title - Nominal Category
overview - Nominal Category
popularity - Measure (ratio)
production_companies - JSON List
production_countries - JSON List
release_date - Ordinal Category
revenue - Measure (ratio)
runtime - Measure (ratio)
spoken_languages - JSON List
status - Nominal Category
tagline - Nominal Category
title - Nominal Category
vote_average - Measure (ratio)
vote_count - Measure (ratio)

Initial Observations

- Lots of columns are currently formatted as lists; must be engineered into meaningful features
  - Cast has actors and the characters they play; crew has all of the writers producers etc. and their names as well. (per movie)
- Only 35.6% (1,712) movies have associated homepages
- 2 movies lack a runtime, 1 movie lackes a release date
- 82.4% (3,959) movies have a tagline
- release_date is an object; should be a datetime
- No duplicates in either

Initial Observations (2)

- The “keywords” feature contains a varying number of words/phrases associated with the genre/style of the movie
- Overview is just a plot synopsis
- Each production company and country has their own name and id, as well as languages
- Vote_average is between 1 and 10.
- Movies each have a different amount of assigned genres

Data Preparation 1

- Before doing any analysis, want to:
  - Combine tables
  - Expand JSON Lists into meaningful features
- Action taken (Crew)
  - Turned ‘cast’ into top_5_actors, consisting of the first five names appearing in the JSON
  - Turned ‘crew’ into Director (containing the director)
    - Other positions like producer or writer sometimes overlapped with director so it was decided to simplify to just director
  - Turned top_5_actors into 5 separate features containing actor names

Data Preparation 1 (2)

- Converted all JSON features in the movies table into comma separated lists
- Merged credits and movies using a left join on movie_id
- Turned genres into genres_1,2, and 3 taking the first three genres in the list
  - Also converted spoken_languages, production_companies and production_countries to the same three feature format
- Dropped keywords for large variance in the number of provided keywords and overlap with genres
  - Some movies had 10 keywords while some had 2; sometimes the keywords were almost exactly the same as the genre and sometimes they were widely different.

Updated Observations

- 99.3% (4,773) movies had directors
- All movies had an actor_1, but only 96.5% (4,635) of movies had an actor_5
- All movies had a genre_1, but 80.6% (3,875) had a genre 2 and 49.7%(2,385) had a genre_3
- Production company 2: 70.6% (3392), company 3: 48.5% (2333)
- Prod country 2: 25.9% (1246), country 3: 8.2% (395)
- Language 2: 28.3% (1362), language 3: 10.8% (521)

- Also changed release_date to datetime here

Univariate Analysis 1

Numeric features summary statistics:

	count	mean	variance	min	max	skewness	kurtosis
budget	4803.0	2.904504e+07	1.658313e+15	0.0	3.800000e+08	2.436450	7.648841
popularity	4803.0	2.149230e+01	1.012299e+03	0.0	8.755813e+02	9.718380	191.794759
revenue	4803.0	8.226064e+07	2.652244e+16	0.0	2.787965e+09	4.443328	33.087909
runtime	4801.0	1.068759e+02	5.112996e+02	0.0	3.380000e+02	0.715733	8.924896
vote_average	4803.0	6.092172e+00	1.427098e+00	0.0	1.000000e+01	-1.959098	7.783004
vote_count	4803.0	6.902180e+02	1.524202e+06	0.0	1.375200e+04	3.822874	19.891973

Univariate Analysis 1 (2)

Other feature Summary Statistics:

	homepage	original_language	original_title	overview	status	tagline	title	director	actor_1	actor_2	actor_3	actor_4
count	1712	4803	4803	4800	4803	3959	4803	4773	4803	4750	4740	4710
unique	1691	37	4801	4800	3	3944	4800	2349	2096	2719	3096	3370
top	http://www.missionimpossible.com/	en	Out of the Blue	In the 22nd century, a paraplegic Marine is di...	Released	Based on a true story.	The Host	Steven Spielberg		Jennifer Aniston	Cameron Diaz	Woody Harrelson
freq	4	4505	2	1	4795	3	2	27	43	15	9	10

	actor_5	genre_1	genre_2	genre_3	production_company_1	production_company_2	production_company_3	production_country_1	production_country_2
count	4635	4803	3875	2385	4803	3392	2333	4803	1246
unique	3546	21	19	20	1311	1613	1443	71	53
top	Steve Buscemi	Drama	Drama	Thriller		Warner Bros.	Warner Bros.	United States of America	United States of America
freq	9	1207	788	424	351	99	86	3102	645

production_country_3	language_1	language_2	language_3
395	4803	1362	521
40	47	49	43
United States of America	English	English	Deutsch
146	4102	323	62

Univariate Analysis 1 (3)

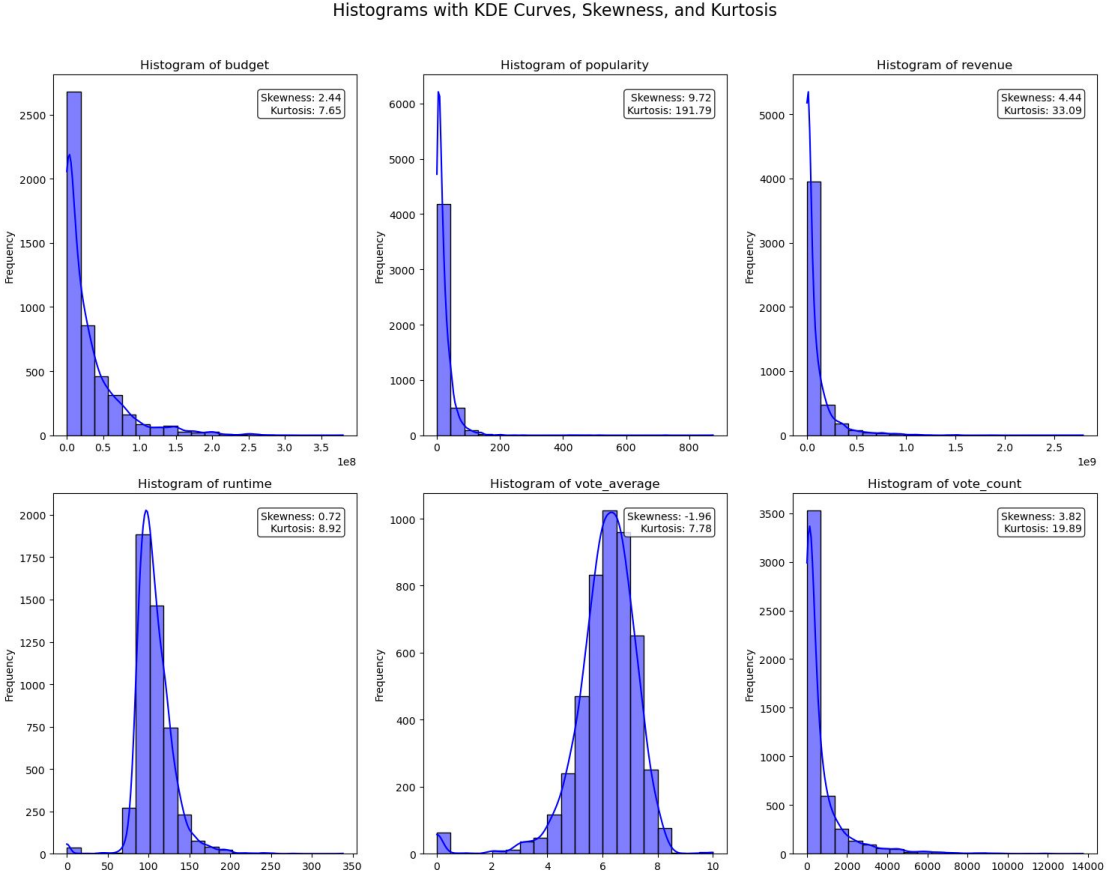
Unique value counts:

```
Unique values in categorical columns:

homepage                1691
original_language         37
original_title          4801
overview                4800
status                     3
tagline                 3944
title                   4800
director                2349
actor_1                 2096
actor_2                 2719
actor_3                 3096
actor_4                 3370
actor_5                 3546
genre_1                   21
genre_2                   19
genre_3                   20
production_company_1    1311
production_company_2    1613
production_company_3    1443
production_country_1      71
production_country_2      53
production_country_3      40
language_1                47
language_2                49
language_3                43
```

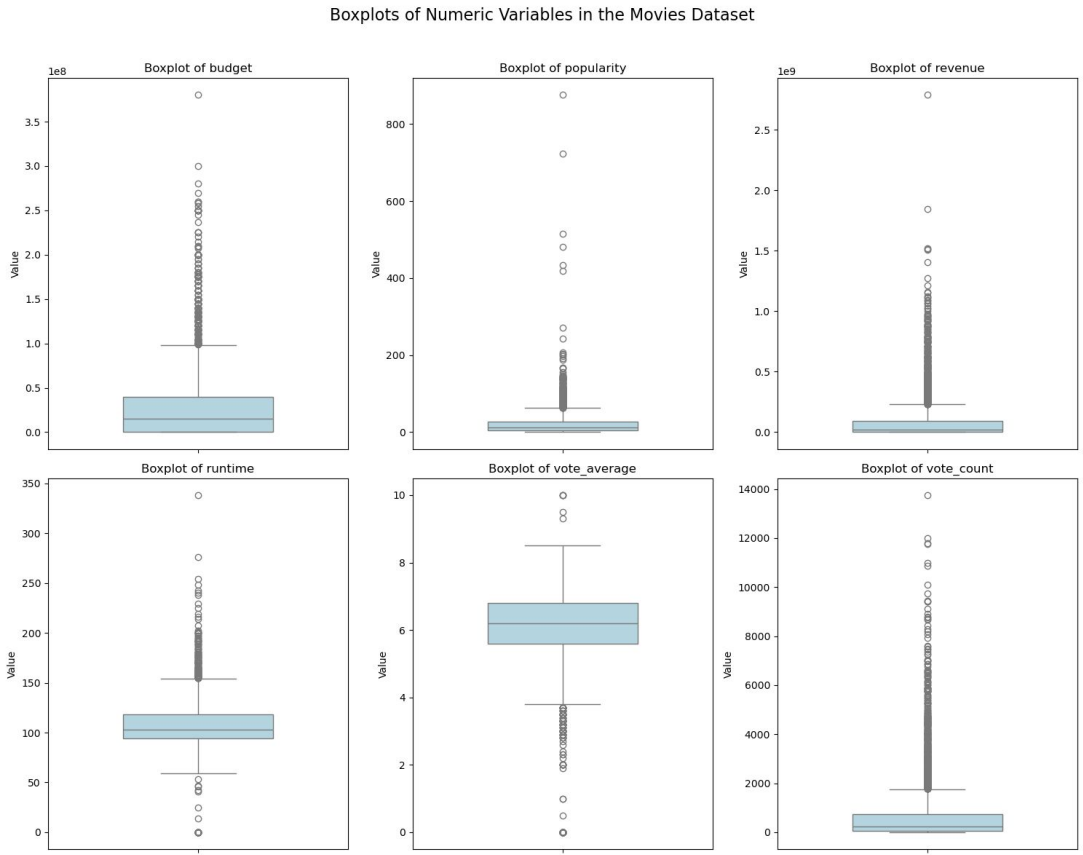
Univariate Visualizations 1

Histograms:



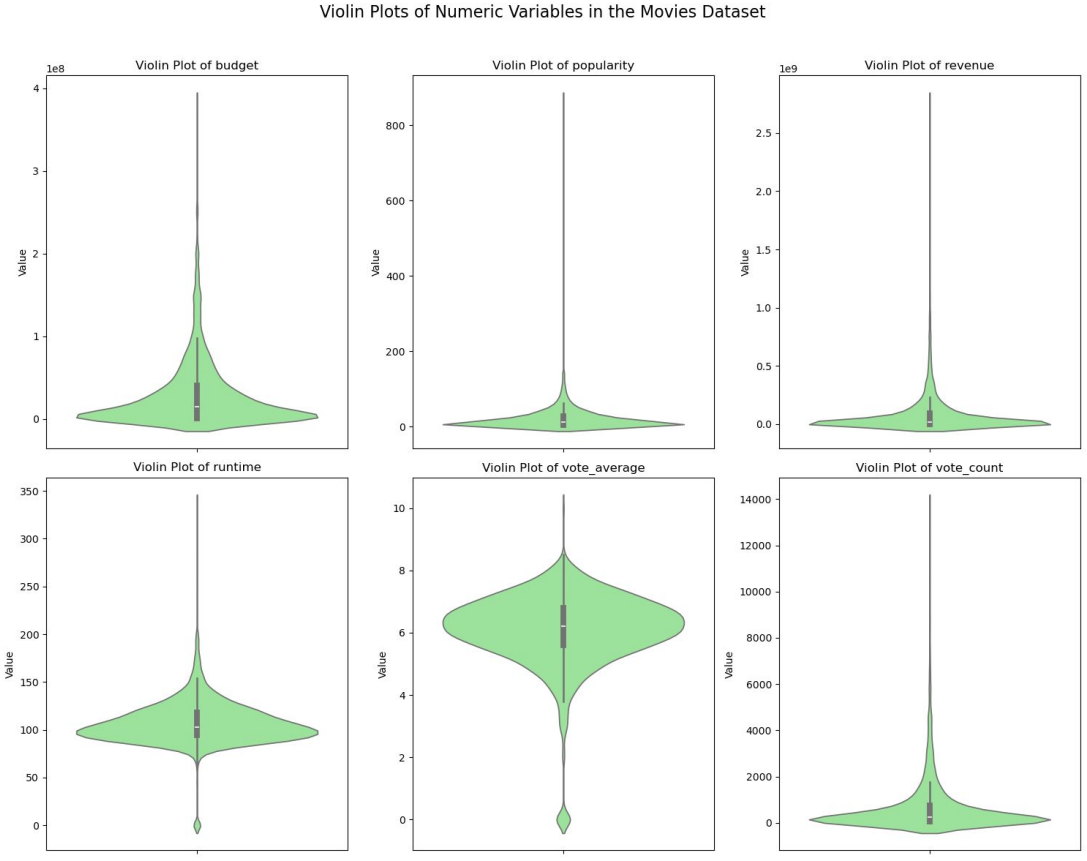
Univariate Visualizations 1 (2)

Box Plots:



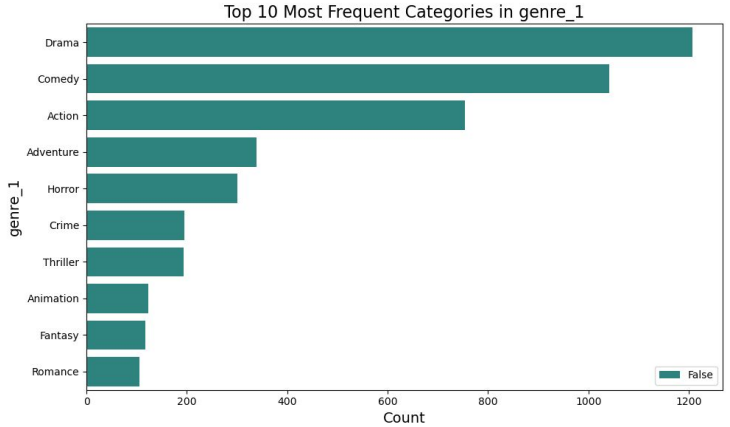
Univariate Visualizations 1 (3)

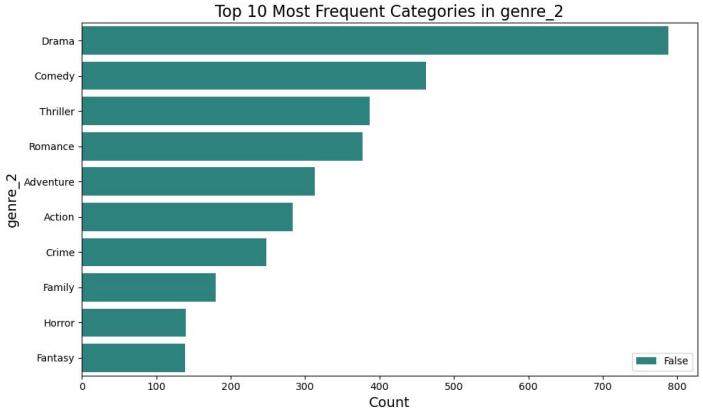
Violin Plots

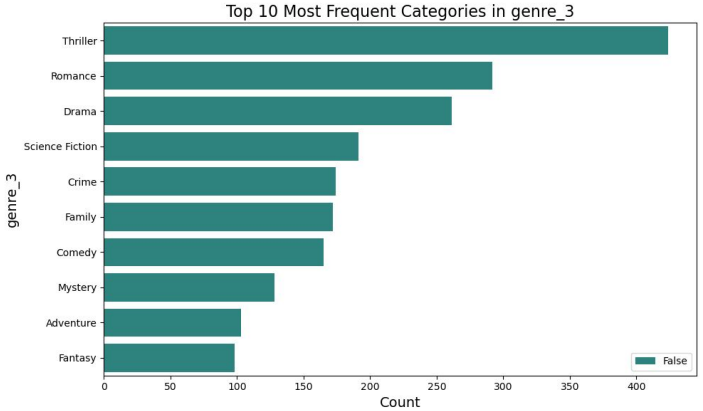


Univariate Visualizations 1 (4)

Genre bar charts:







Data Preparation 2

- At this point, I realized I would need to refine the features even more; specifically the ones with multiples of itself
  - The graphs often featured overlaps as seen in genre and caused confusion
  - Some features also had extreme amounts of missing/no values
  - Figured it would make more sense to fix this first before doing further/more complex analysis

Data Preparation 2 (2)

Data Shaping

- Columns dropped: ‘homepage’, ‘tagline’, ‘overview’
  - Not useful towards objectives and had lots of variance/possible values
- Also dropped: x_2 and x_3 of: language, production_company, production_country, and genre
  - Either simply had too many missing values and/or no unifying methodology for which/how many genres were assigned to a movie

Data Preparation 2 (3)

Data Shaping:

- Checking movies with missing values
  - Release_date: only 1 movie that could not be found online and featured 0s in every other column. Verdict: removed
  - Runtime: 2 movies that had no recorded revenue and missing actor values 2-5. Verdict: flagged for now
  - Director: 29 movies had no director. Comparing them to their original JSON list almost all had an empty set for cast and/or crew. Verdict: removed
- At this point, there is now 4773 rows x 22 columns

Data Preparation 2 (4)

- At this point the only missing values are within the actors columns (other than flagged rows)
- Considering the confusion with the columns already and the possibility of movies with <5 actors, I decided to reconfigure the features
- New features: num_actors, lead_actor, num_famous_actors, famous_lead
  - Num_actors: total count of actors in the movie
  - Lead_actor: name of the first actor in the list
  - Num_famous_actor: count of famous actors in the movie. Famous actors are actors with >= 20 appearances in the dataset
  - Famous_lead: boolean for if the lead actor is also a famous actor
  - Removed the features actor_1-actor_5

Data Preparation 2 (5)

Feature Engineering Continued:

- I also decided to rework the production_company and country features.
  - New features: production_company_main, multiple_production_companies; production_country_main, multiple_production_countries
    - Respectively contain the main/first company/country and a boolean on if there were multiples involved.
- I also dropped langauge_1 since it felt unnecessary with original_language also existing
- 326 companies did not have a production_company at this point. I decided to flag them for now.

Outlier Checking

- Based off of checking for nulls as well as initial visualizations, I am suspicious of issues in budget, revenue, runtime, and num_actors
  - Budget and revenue had extremely high counts of values on the low end for budget and revenue
  - Runtime had a small bump around zero on its histogram as well
- An initial z-score check for outliers on budget only provided movies with very high budgets that came out recently and had many famous actors, making sense that they are outliers but not harmful.
  - I assumed this was because there were so many movies with incorrect budgets that they skewed the distribution enough to not be traditionally considered outliers.

Outlier Checking (2)

- 21.7% (1,036/4773) of the movies had budgets below 100 dollars.
  - 792 of those movies also didn’t have a production company
- I decided to continue testing for outliers using the dataset without those 1,036 flagged movies
- On top of the 1,036 <100 budget movies, there were 514 different movies with a revenue of zero dollars
  - Comparing some examples to the website often showed discrepancies between them
- When outlier checking runtime and num_actors, I also dropped the zero revenue movies

Outlier Checking (3)

- Z-score checking runtime returned no anomalous behavior, only very long (3+ hour) movies and one very short (40 minute) movie
- Using z-scores for num_actors returned 71 movies; most of these were blockbuster films with huge casts
    - I also checked for movies with < 3 actors, which returned seven rows.
    - After cross referencing with the tMDB website, I removed the rows with no actors as well as the movies that did not exist on the website. There was, however, a movie that only featured one actor, which I thought was funny.

Data Preparation 2 (6)

- I decided to remove all the movies I had flagged for suspicious budget and revenue. Although it was a large chunk, they had inconsistencies in the most important columns and seemed to be a case of missing at completely random.
- The updated dataset had 3211 rows and 22 columns. 37 rows still lacked a production company and 14 had no country.
  - 11 of the 14 with no country also had no company; I decided to remove those that had neither.
  - The three remaining movies with no country had their information on the tMDB website, so I decided to manually fill them in.
  - The 26 movies with no company seemed to be cases of MCAR, meaning they were safe to remove.

Univariate Analysis 2

The dataset, now of size 3174 x 22, was ready for continued analysis.

Numeric summary statistics:

	budget	popularity	release_date	revenue	runtime	avg_user_score	num_votes	num_actors	num_famous_actors	release_year
count	3.174000e+03	3174.000000	3174	3.174000e+03	3174.000000	3174.000000	3174.000000	3174.000000	3174.000000	3174.000000
mean	4.129071e+07	29.475862	2002-04-01 19:41:51.153119104	1.232553e+08	110.897290	6.312161	993.340895	26.281664	2.204159	2001.712350
min	7.000000e+03	0.037073	1916-09-04 00:00:00	7.000000e+00	41.000000	0.000000	0.000000	1.000000	0.000000	1916.000000
25%	1.100000e+07	10.926848	1998-08-20 06:00:00	1.786998e+07	97.000000	5.800000	190.000000	15.000000	1.000000	1998.000000
50%	2.600000e+07	20.801659	2005-08-04 12:00:00	5.722771e+07	107.000000	6.300000	485.000000	19.000000	2.000000	2005.000000
75%	5.500000e+07	37.764689	2010-11-29 06:00:00	1.491741e+08	121.000000	6.900000	1168.750000	30.000000	3.000000	2010.000000
max	3.800000e+08	875.581305	2016-09-09 00:00:00	2.787965e+09	338.000000	8.500000	13752.000000	224.000000	15.000000	2016.000000
std	4.450746e+07	36.314834	NaN	1.872728e+08	20.966473	0.869925	1421.112279	21.558386	2.038837	13.243916

Univariate Analysis 2 (2)

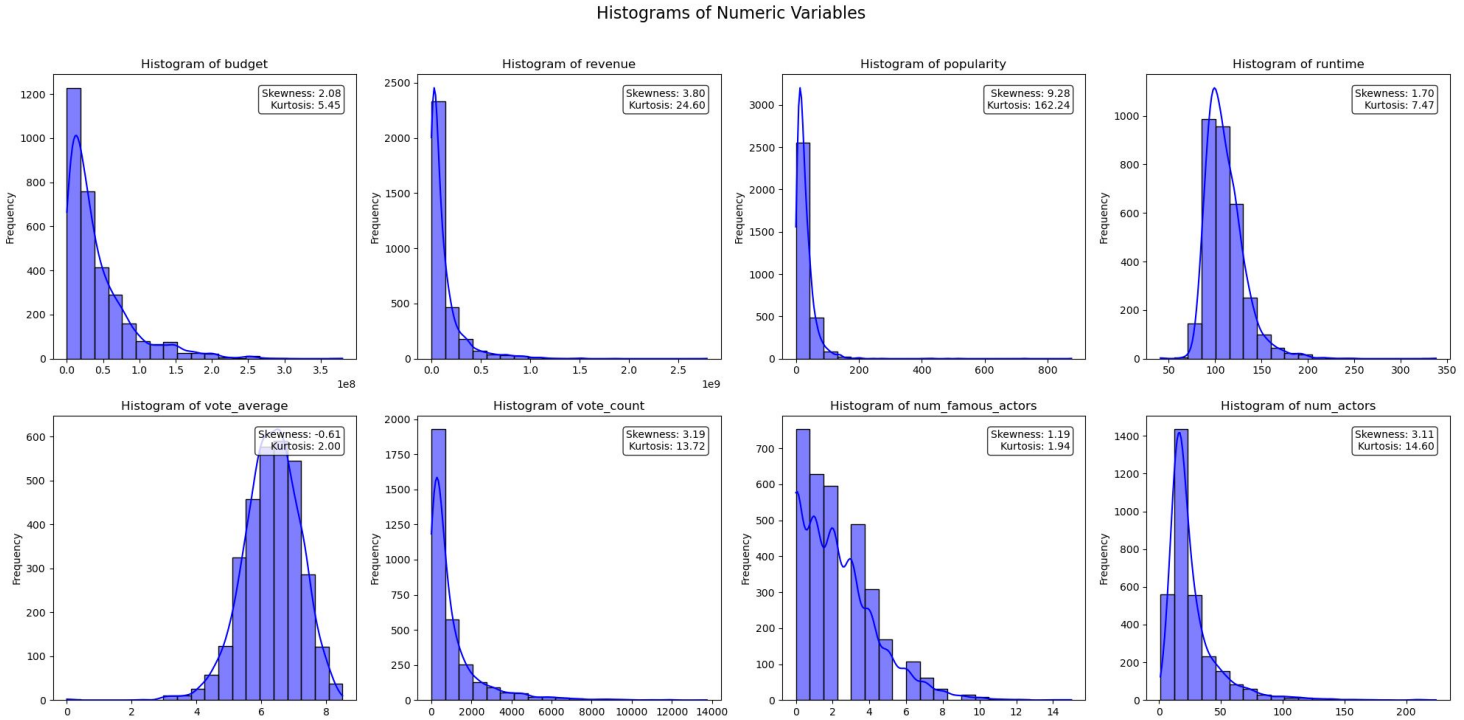
Unique counts:

```
Unique values in categorical columns after cleaning:

original_language           25
original_title            3174
title                     3173
director                  1410
genre                       19
lead_actor                1266
production_company_main    748
production_country_main     45
```

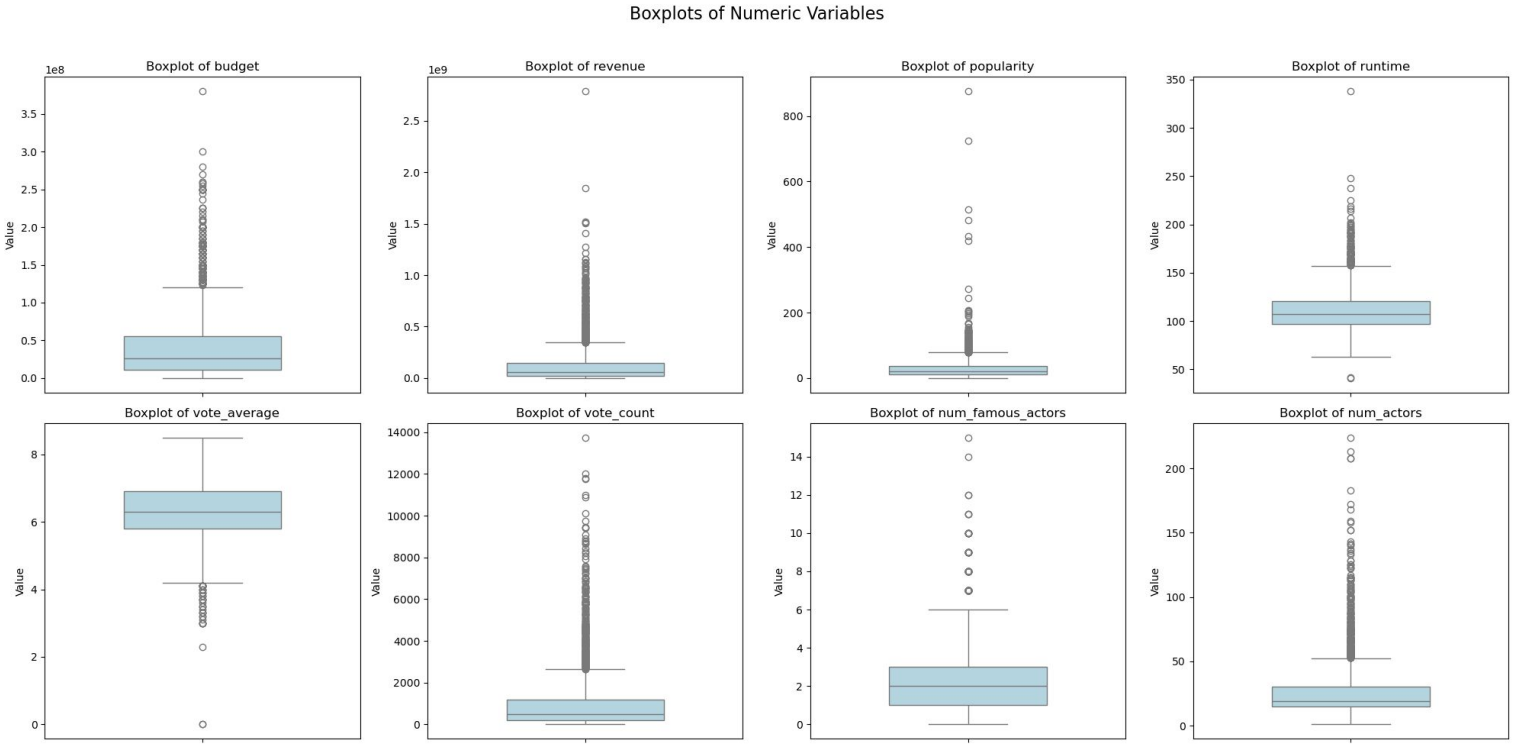
Univariate Analysis 2 (3)

Histograms:



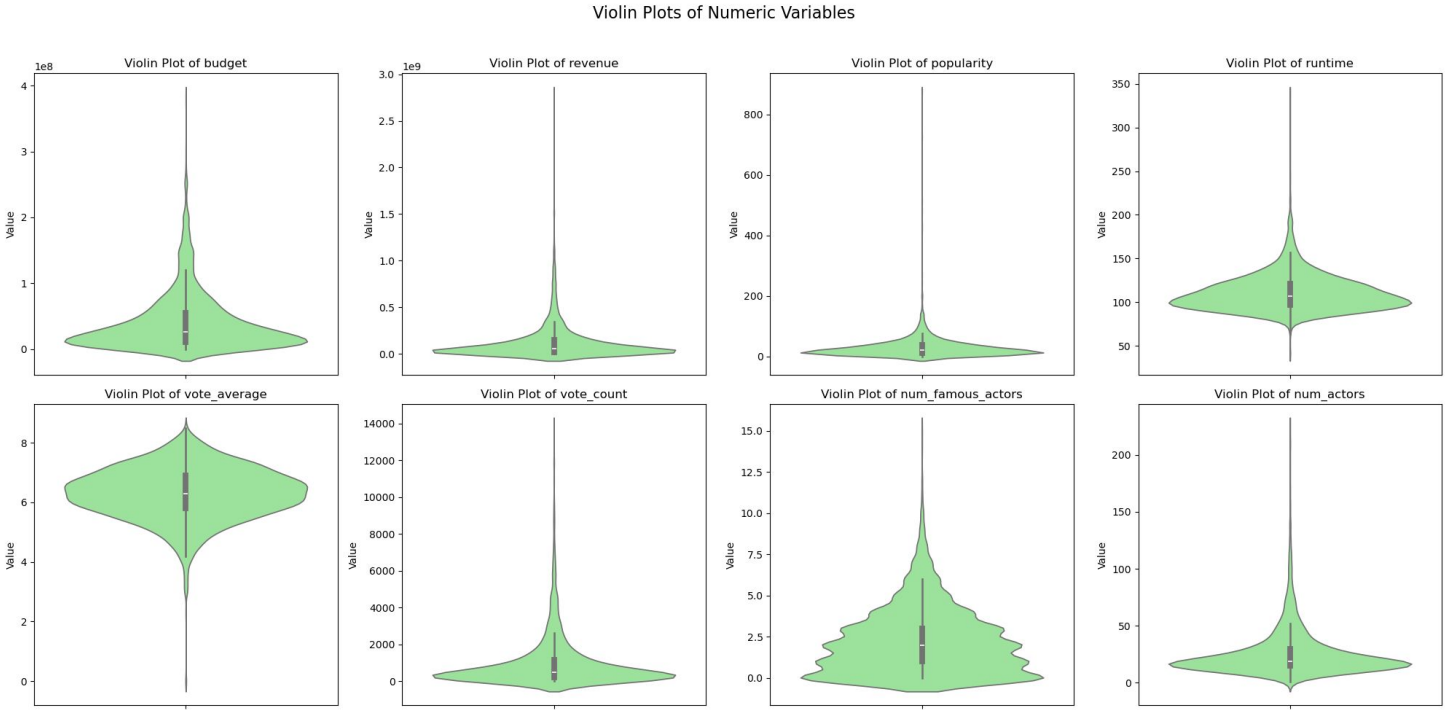
Univariate Analysis 2 (4)

Box plots:



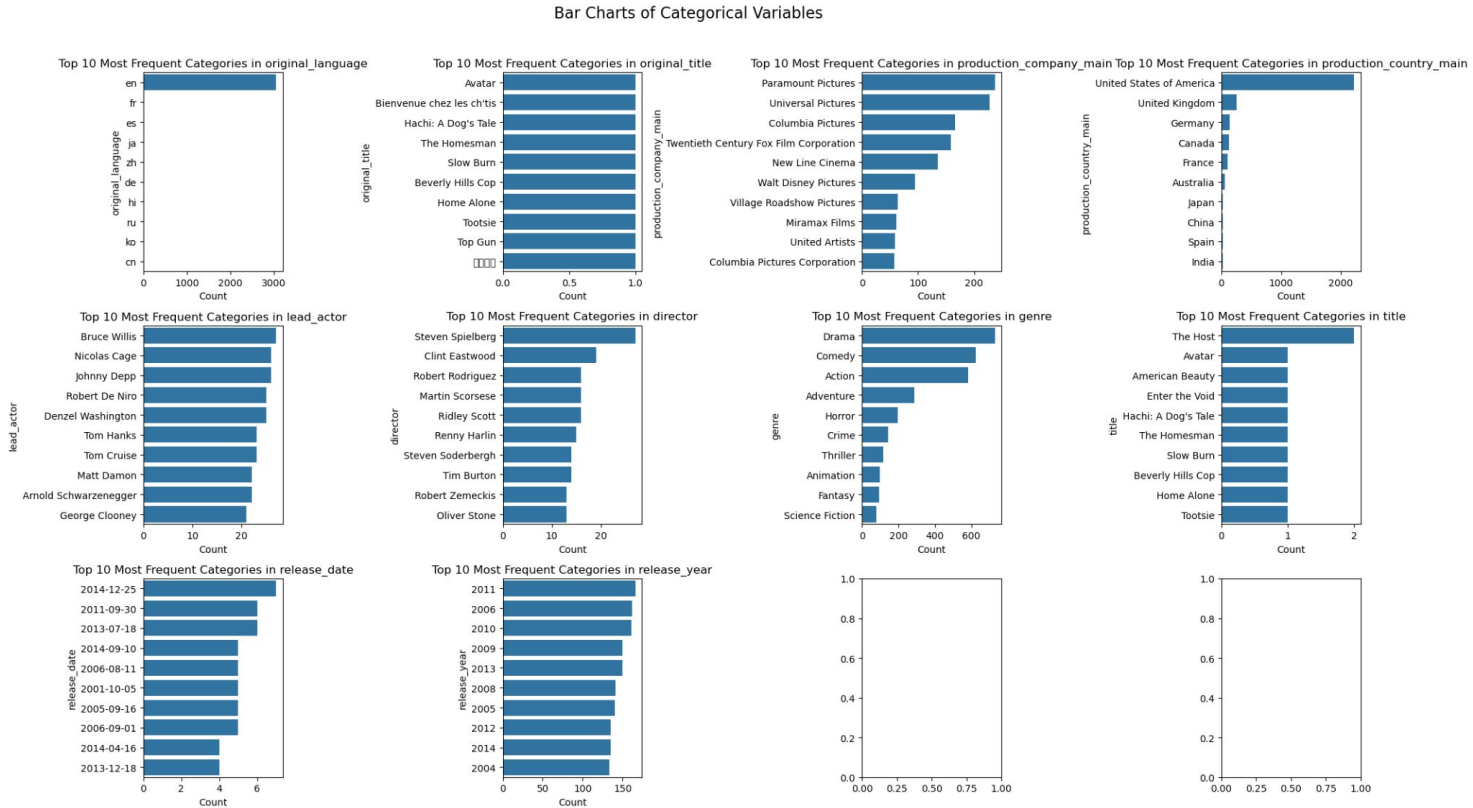
Univariate Analysis 2 (5)

Violin Plots



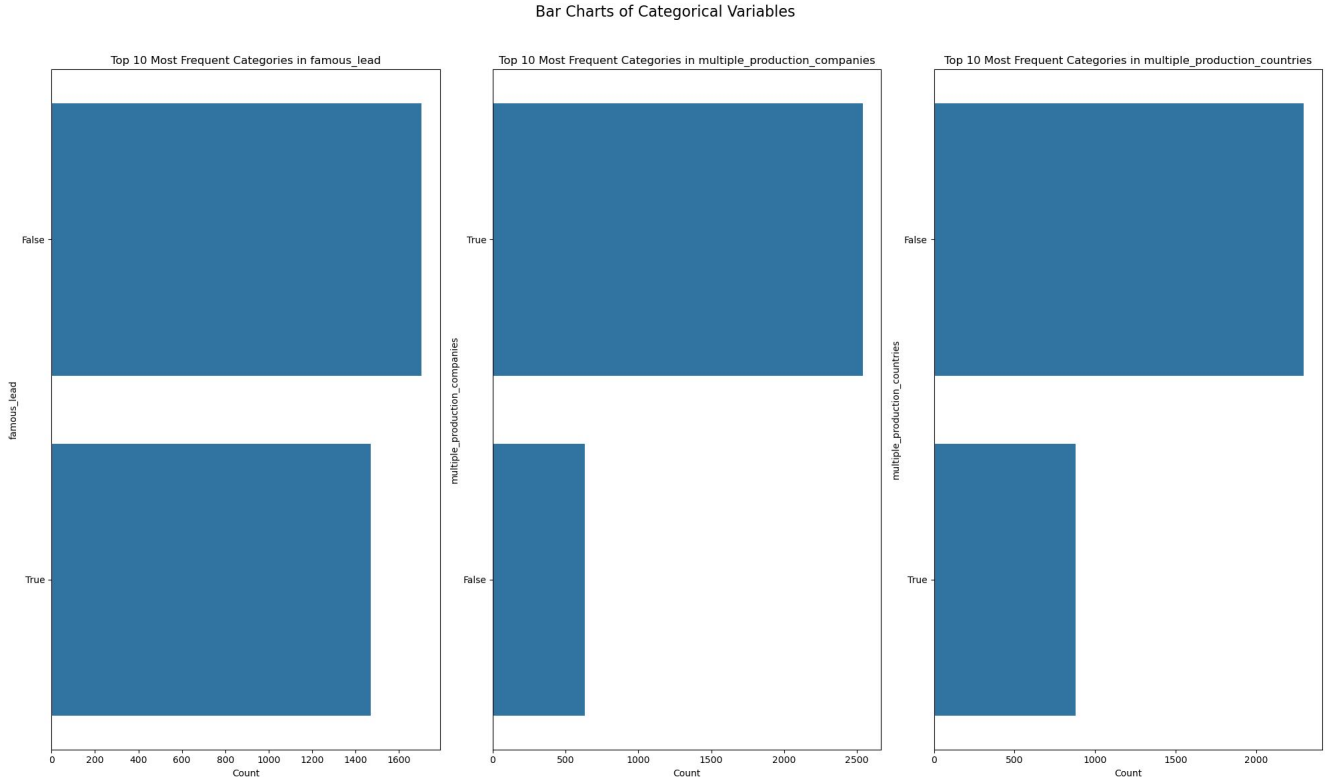
Univariate Analysis 2 (6)

Bar Charts
(Top 10 only)



Univariate Analysis 2 (7)

Boolean feature bar charts:

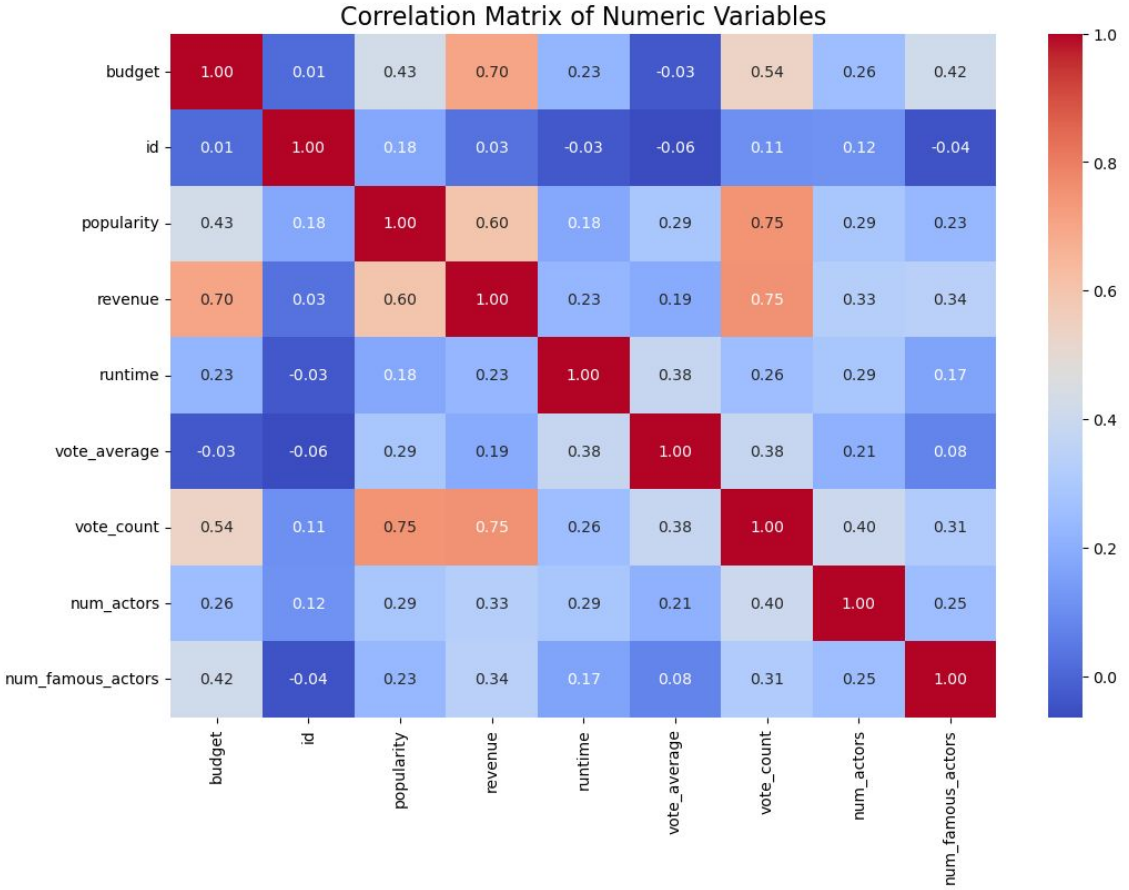


Univariate Analysis 2 (8)

- The distributions align with my domain knowledge and make sense to me
- Budget is more inline to what I expected compared to pre-cleaning visualizations
- Original language is dominated by english
- Some features are a bit difficult to display as they have a large-to-huge amount of distinct values

Bivariate Analysis

Feature Correlation Matrix:



Bivariate Analysis (2)

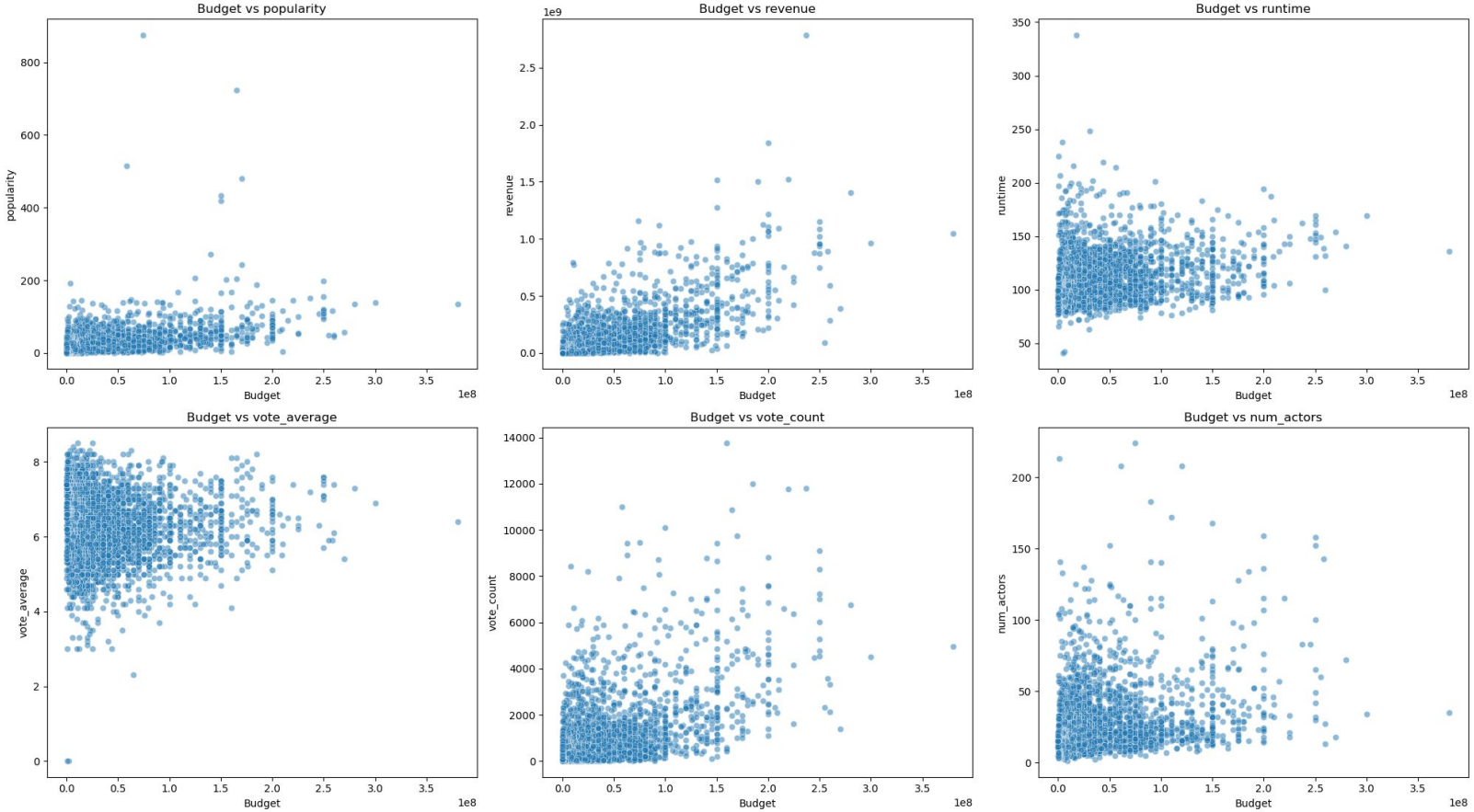
- Most correlations make sense and align with my domain knowledge
- Budget is correlated with revenue and popularity which makes sense and is not correlated enough to be problematic
    - Vote_count is also a bit less correlated with budget + revenue and more so with popularity. It could be a case of multicollinearity between them (vote_count and popularity) but further testing is required I believe.

Bivariate Analysis (3)

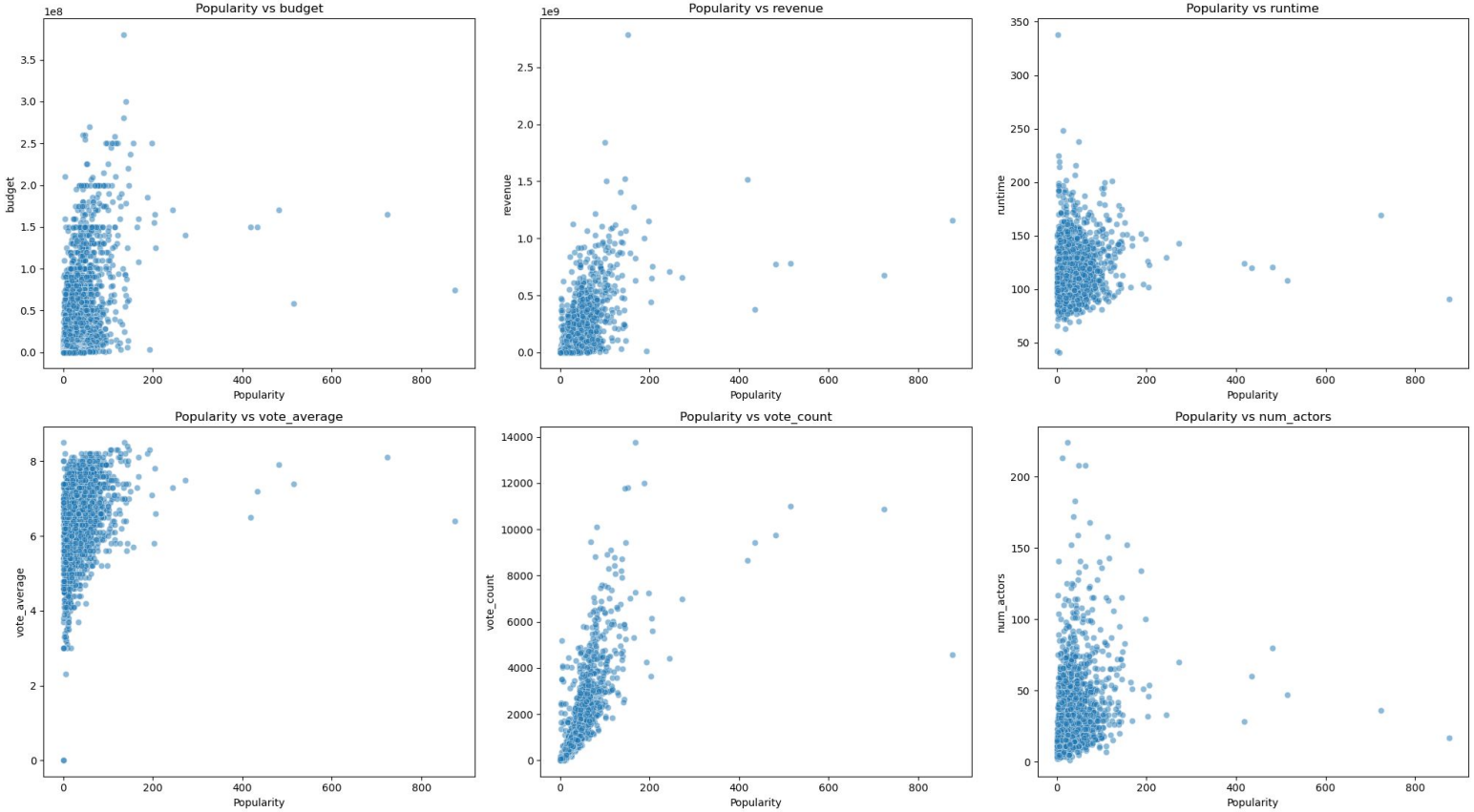
The following slides will be all the permutations of scatterplots of the numeric variables plotted against each other.

- From my understanding, everything looks to be within the expected range.
- Num_famous_actors should maybe be not treated like a numeric

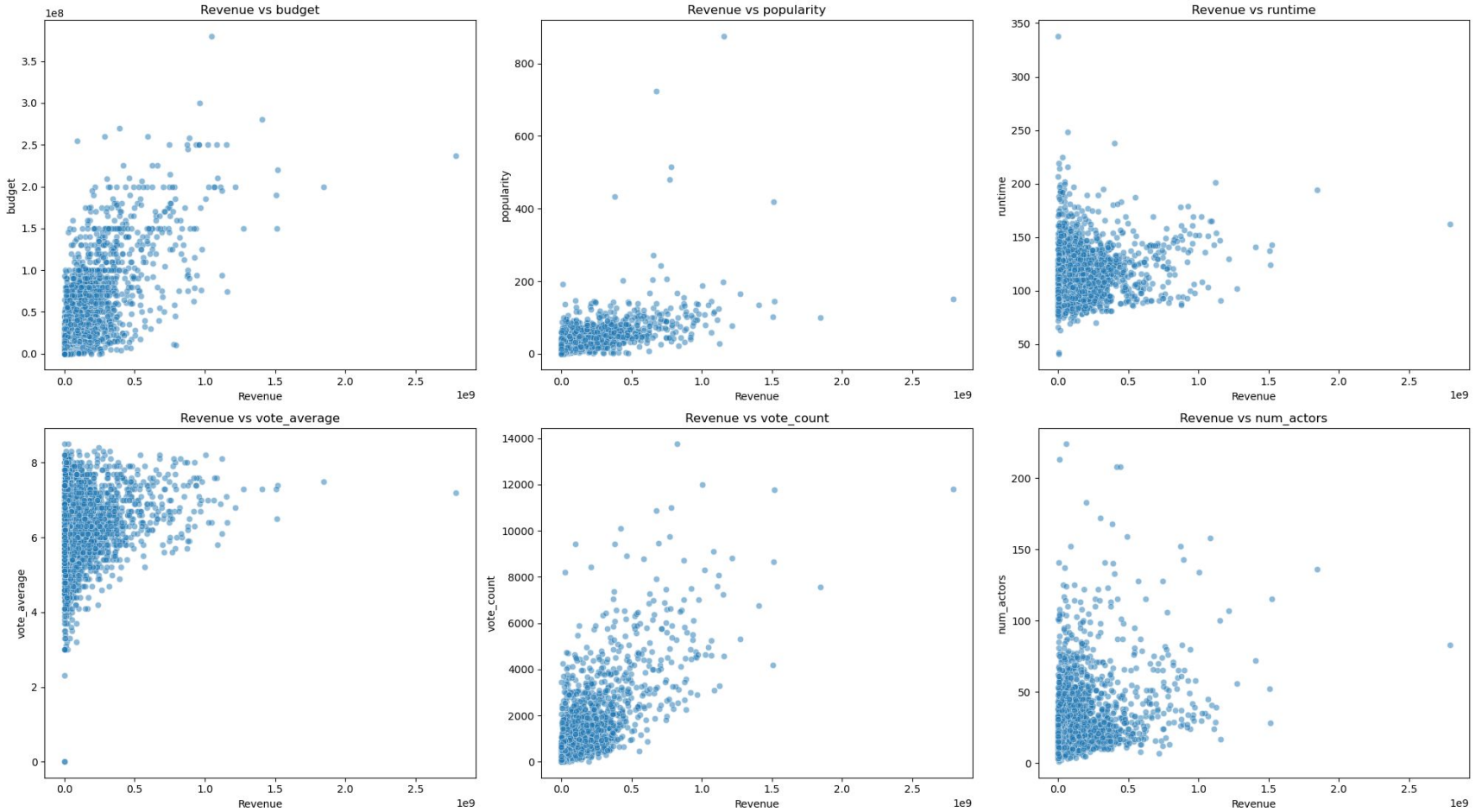
Scatter Plots of Budget vs Other Numeric Variables



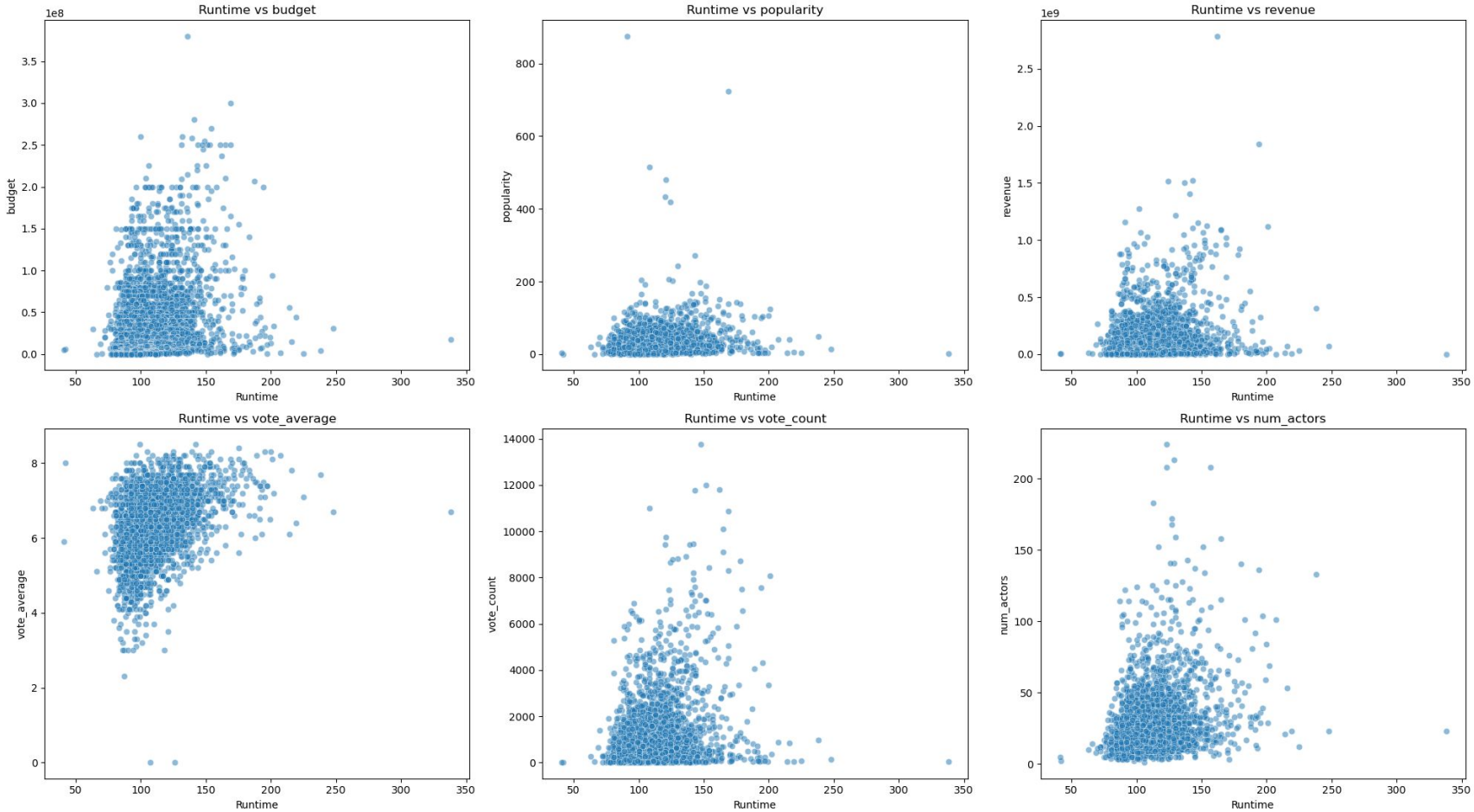
Scatter Plots of Popularity vs Other Numeric Variables

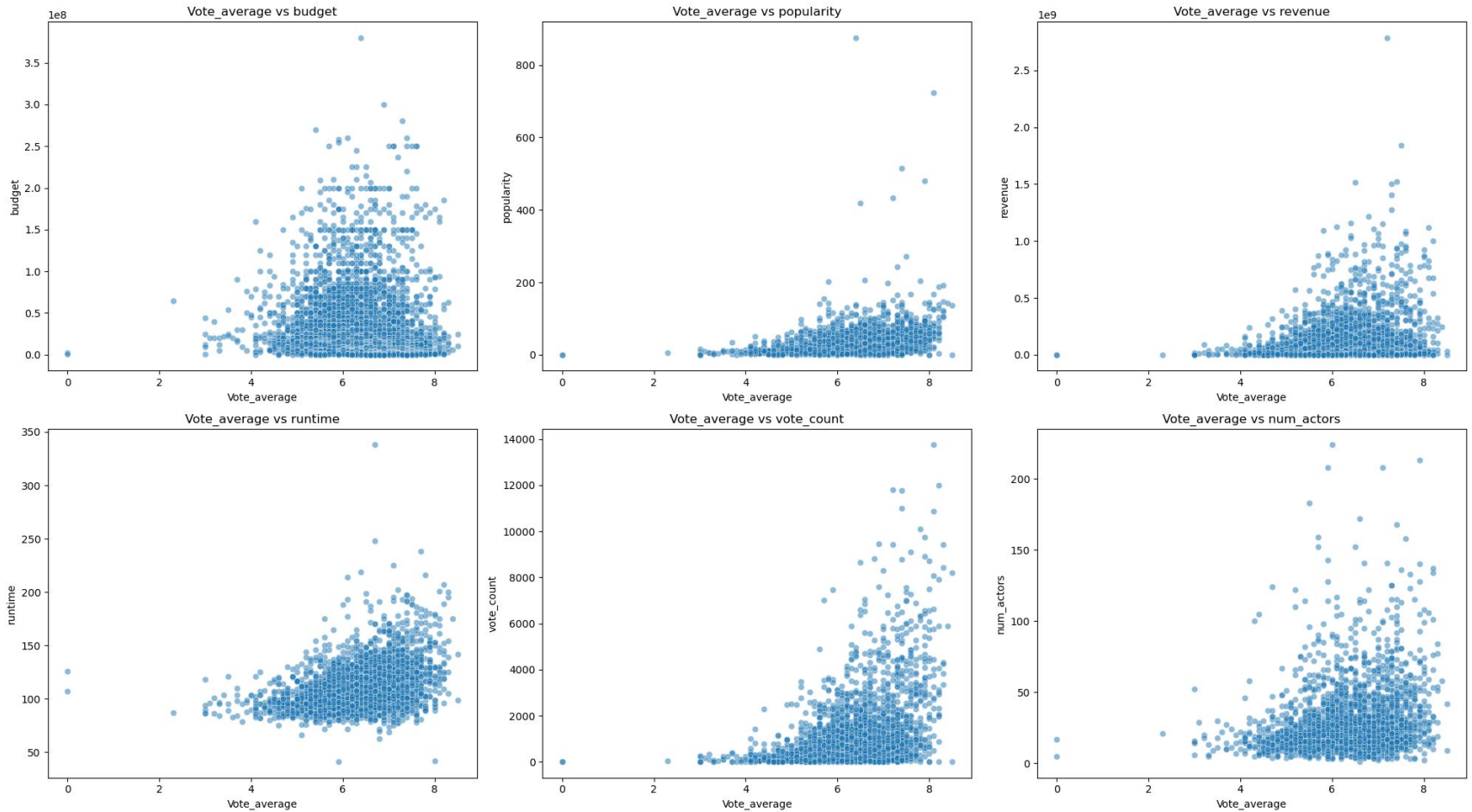


Scatter Plots of Revenue vs Other Numeric Variables

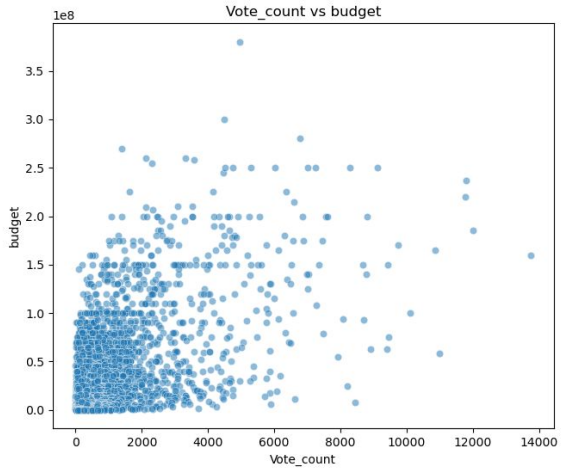


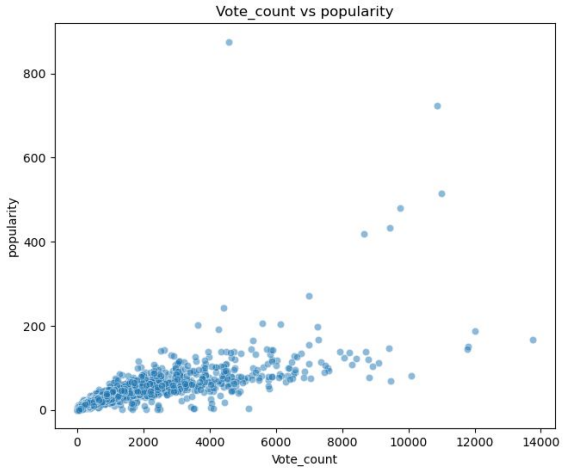
Scatter Plots of Runtime vs Other Numeric Variables

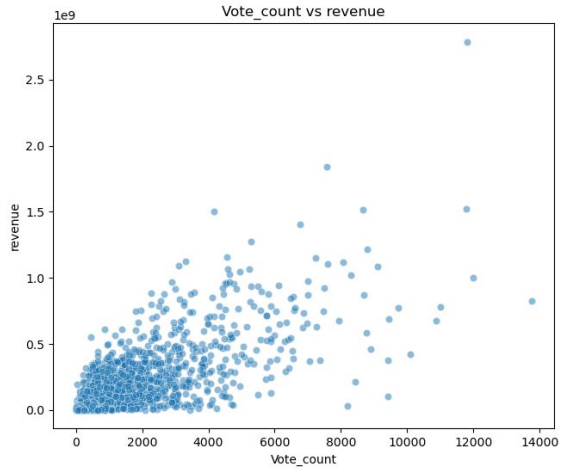


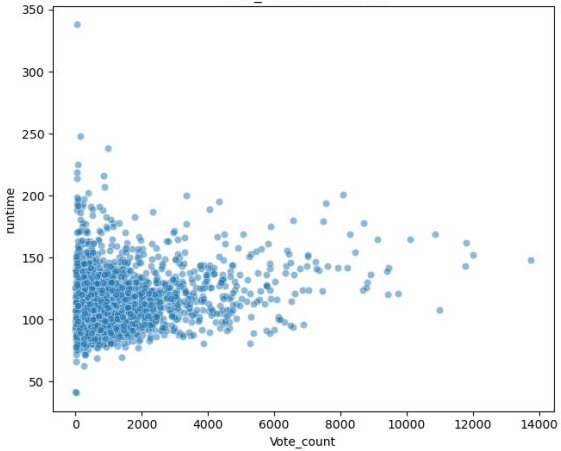

Scatter Plots of Vote_average vs Other Numeric Variables
Vote_average vs budget
1e8
budget
Vote_average
Vote_average vs popularity
popularity
Vote_average
Vote_average vs revenue
1e9
revenue
Vote_average
Vote_average vs runtime
runtime
Vote_average
Vote_average vs vote_count
vote_count
Vote_average
Vote_average vs num_actors
num_actors
Vote_average

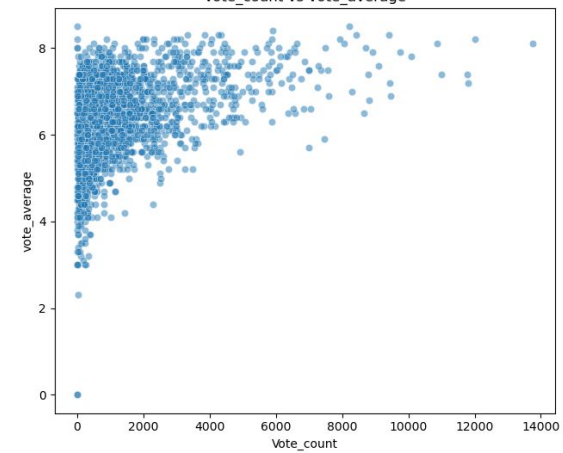
Scatter Plots of Vote_count vs Other Numeric Variables

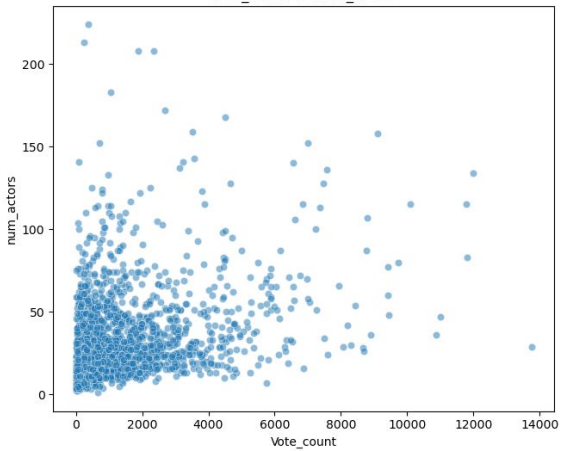


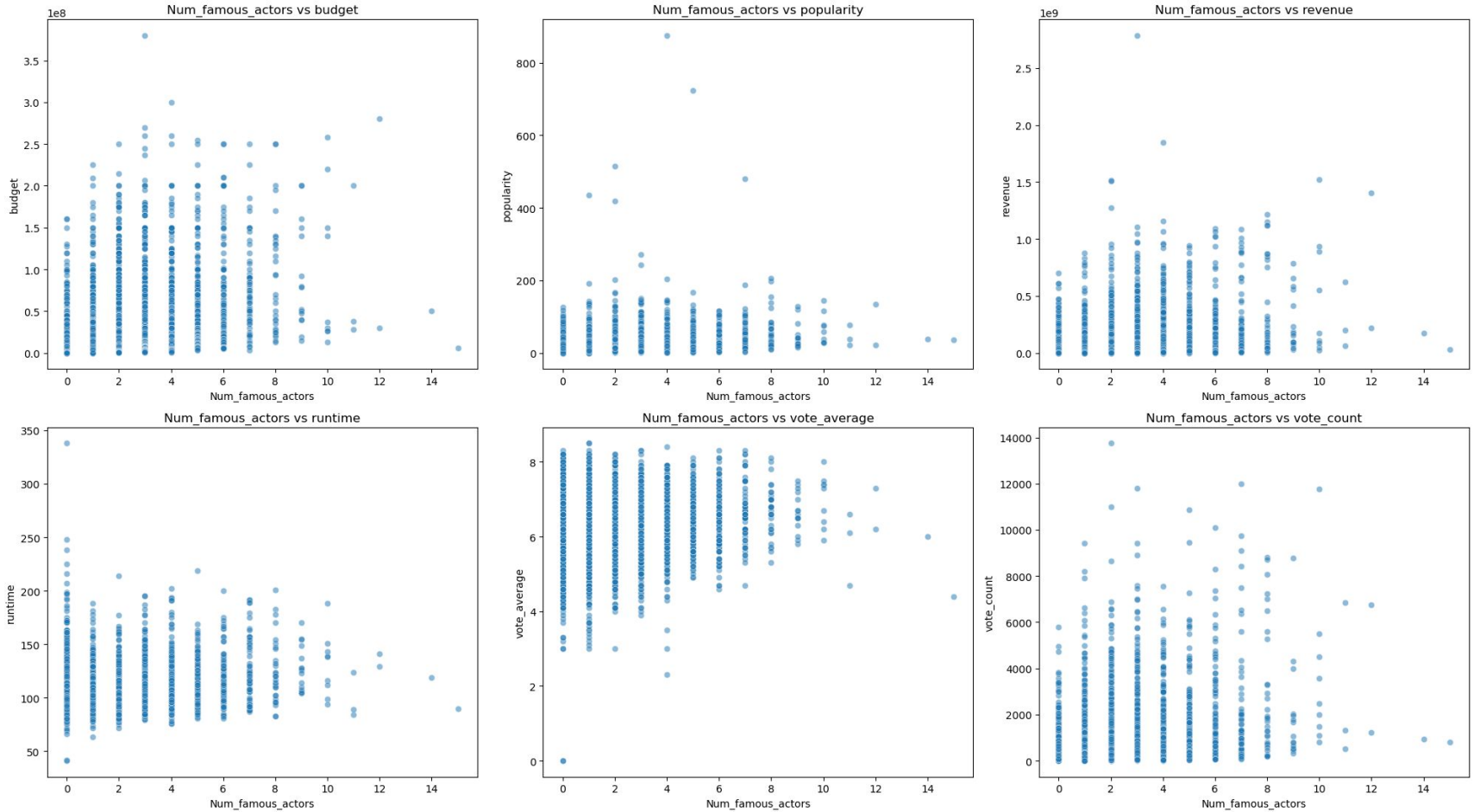










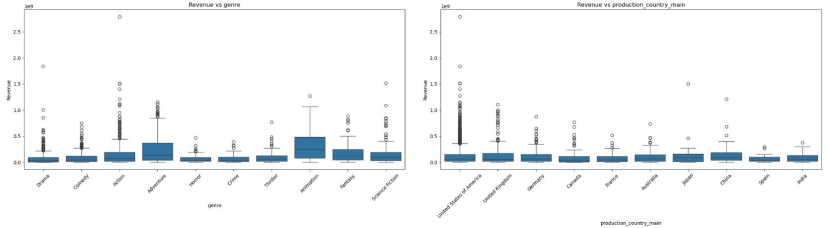

Scatter Plots of Num_famous_actors vs Other Numeric Variables
Num_famous_actors vs budget
1e8
budget
Num_famous_actors
Num_famous_actors vs popularity
popularity
Num_famous_actors
Num_famous_actors vs revenue
1e9
revenue
Num_famous_actors
Num_famous_actors vs runtime
runtime
Num_famous_actors
Num_famous_actors vs vote_average
vote_average
Num_famous_actors
Num_famous_actors vs vote_count
vote_count
Num_famous_actors

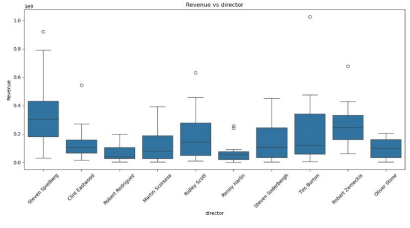
Bivariate Analysis (4)

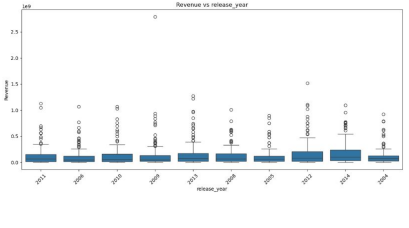
The following slides will be multiple box plots for the top 10 most common values of the categorical features.

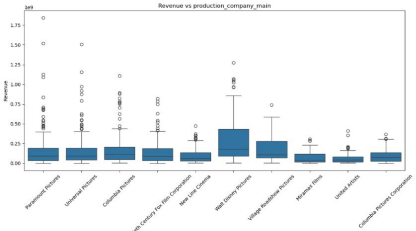
- More insightful to the questions trying to be answered rather than outliers/errors in the dataset to be hoenst

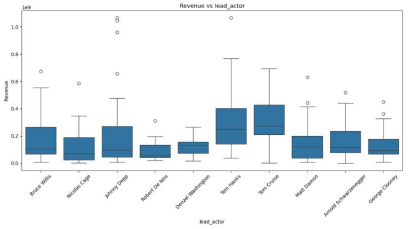
Box Plots of Revenue vs Categorical Variables

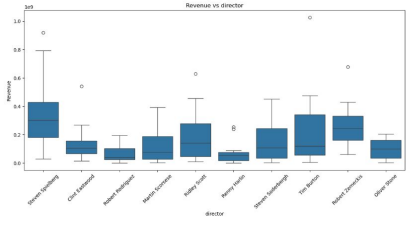


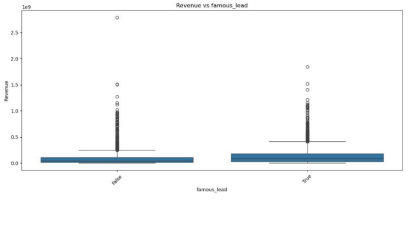




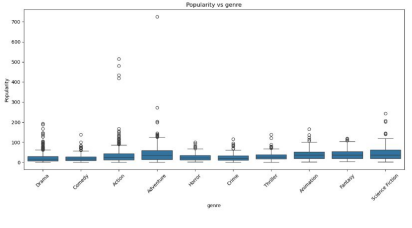


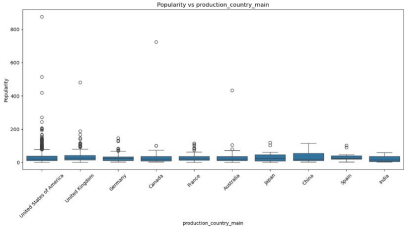


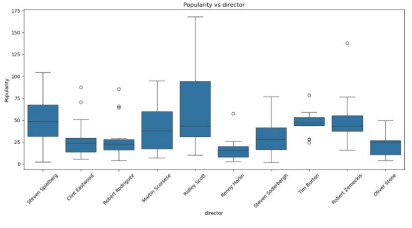


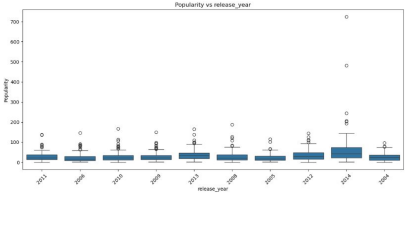


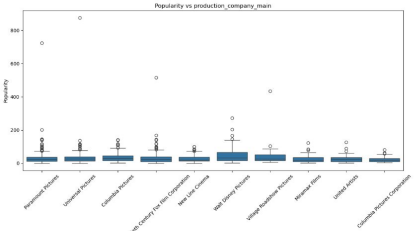
Box Plots of Popularity vs Categorical Variables

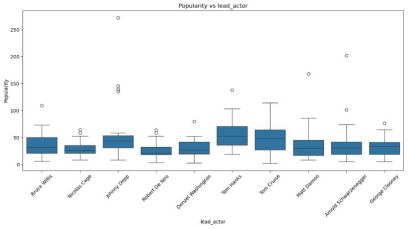


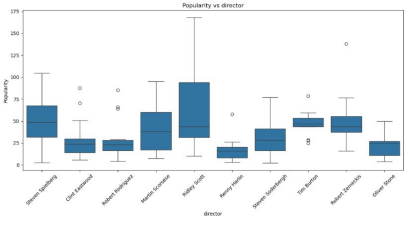


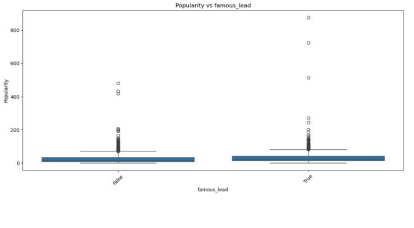




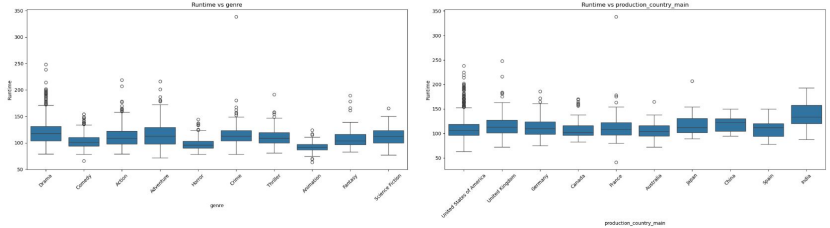


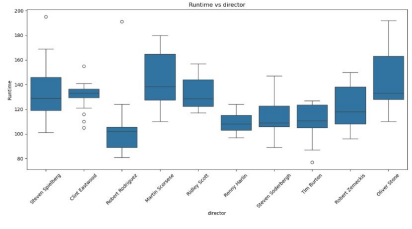


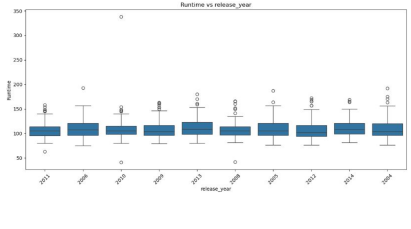


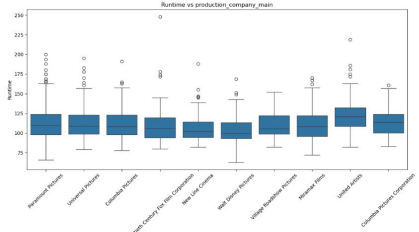


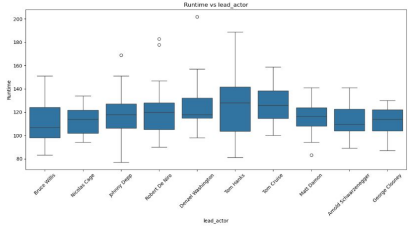
Box Plots of Runtime vs Categorical Variables

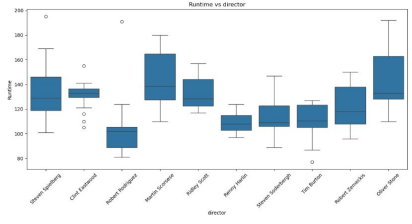


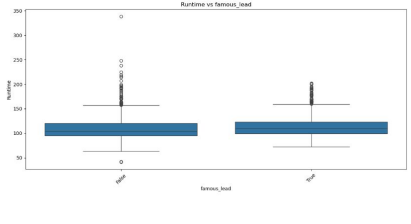




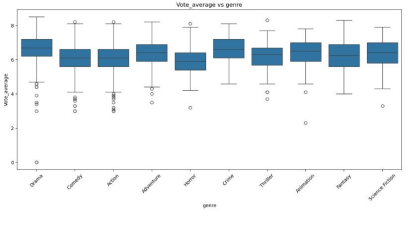


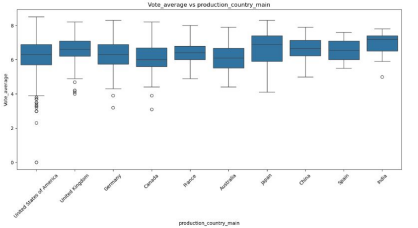


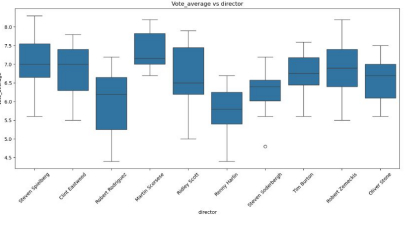


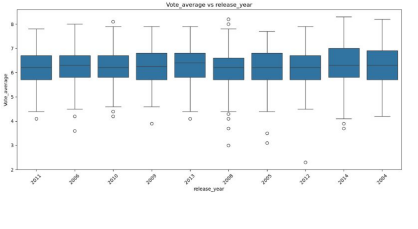


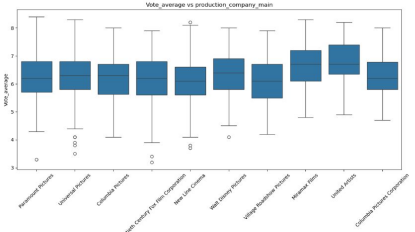
Box Plots of Vote_average vs Categorical Variables

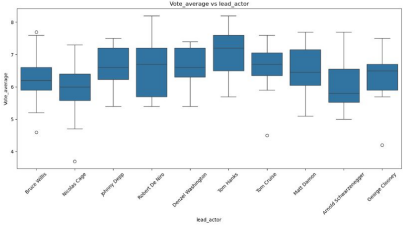


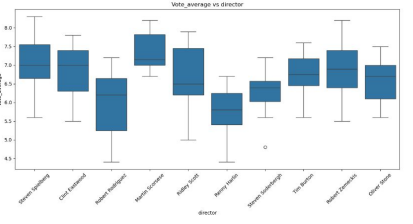


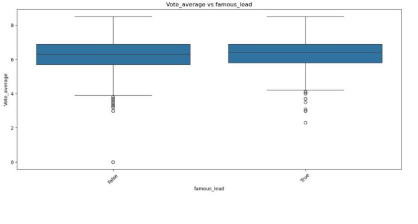




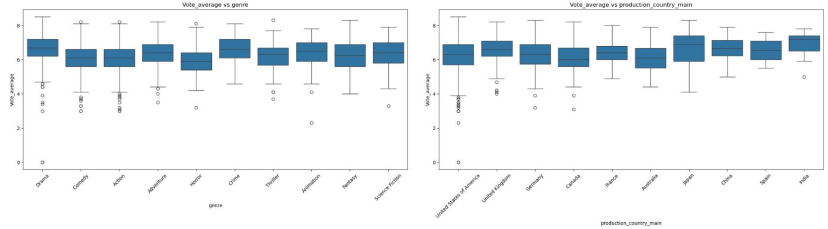


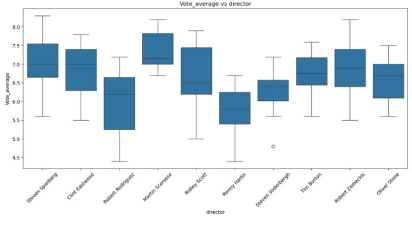


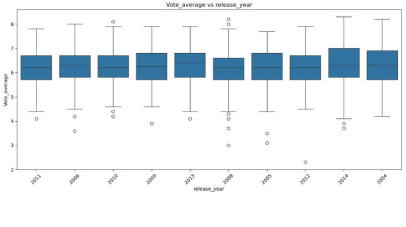


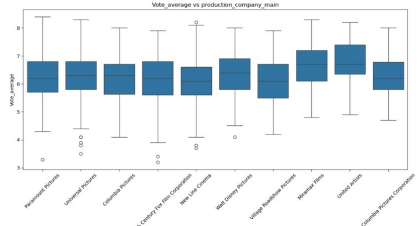


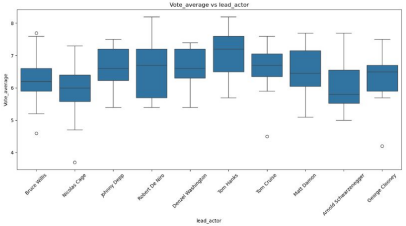
Box Plots of Vote_average vs Categorical Variables

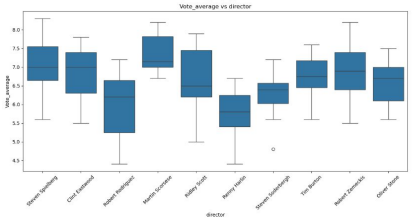


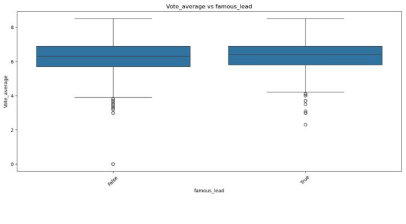




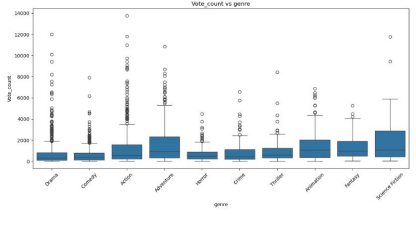


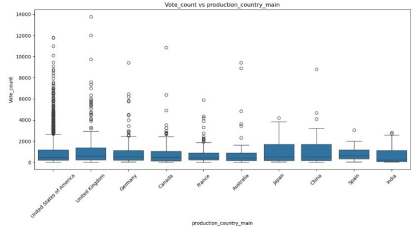


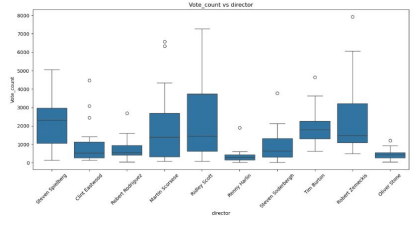


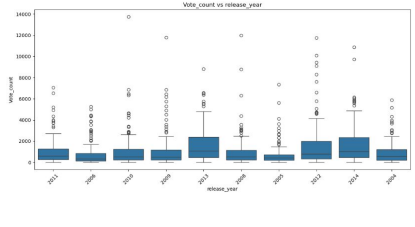


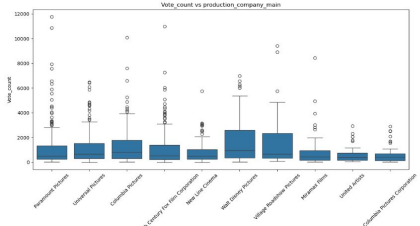
Box Plots of Vote_count vs Categorical Variables

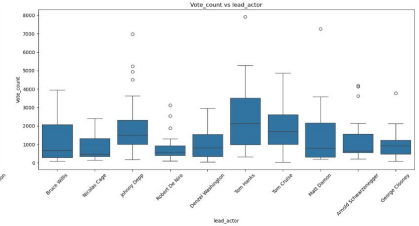


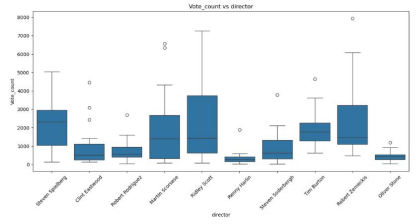


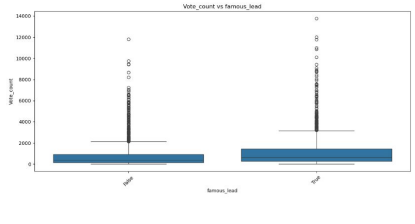




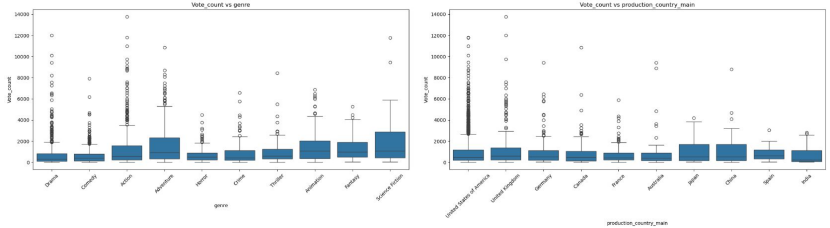


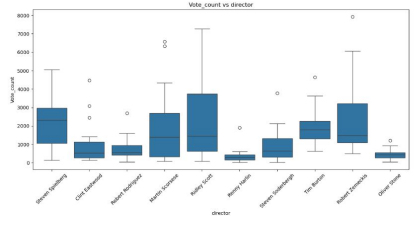


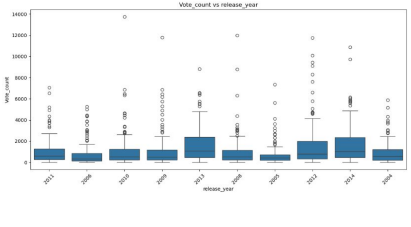


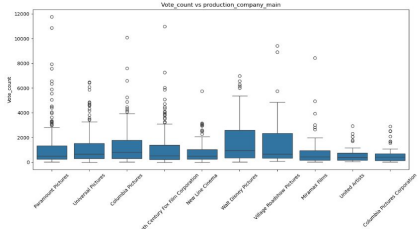


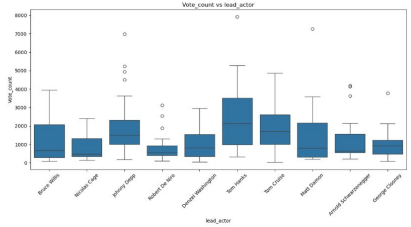
Box Plots of Vote_count vs Categorical Variables

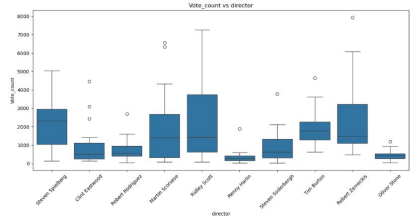


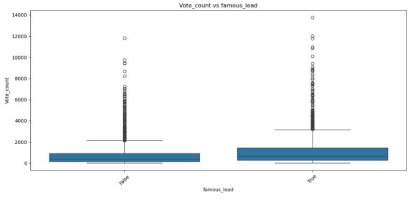




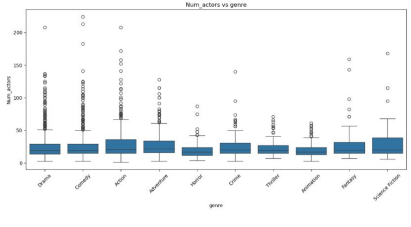


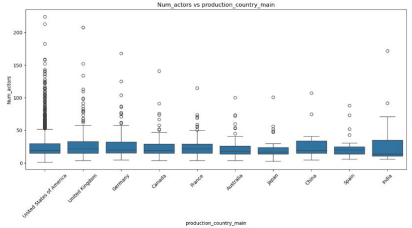


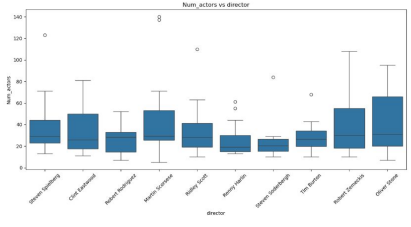


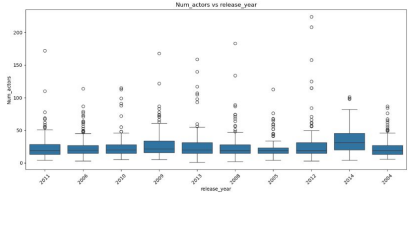


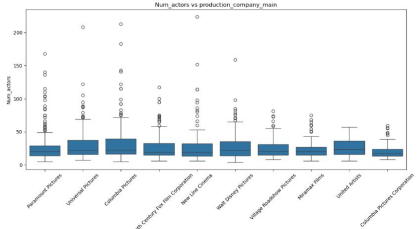
Box Plots of Num_actors vs Categorical Variables

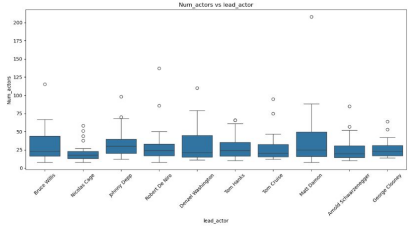


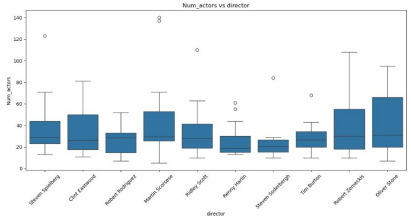


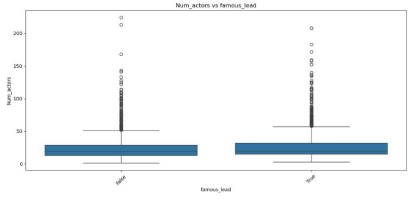




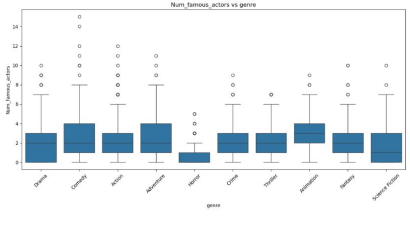


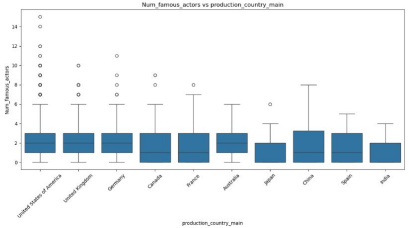


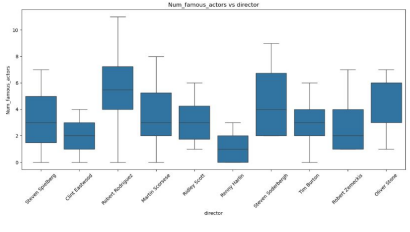


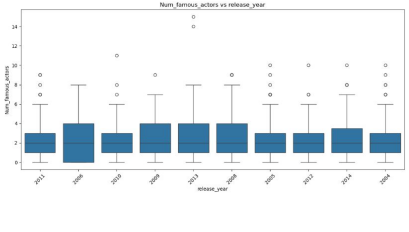


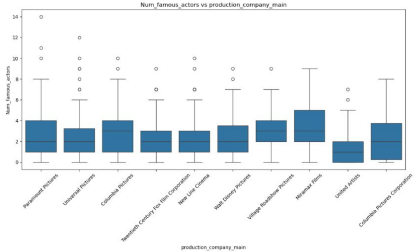
Box Plots of Num_famous_actors vs Categorical Variables

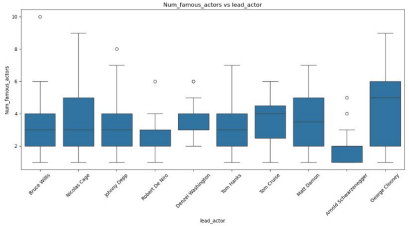


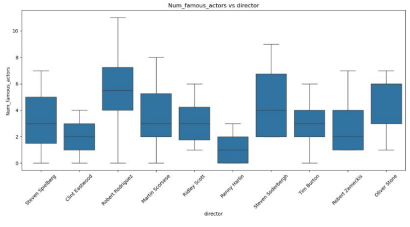


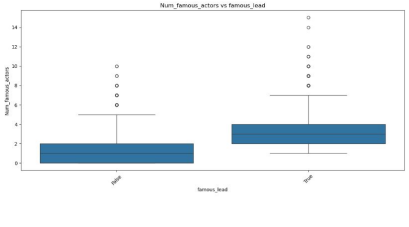






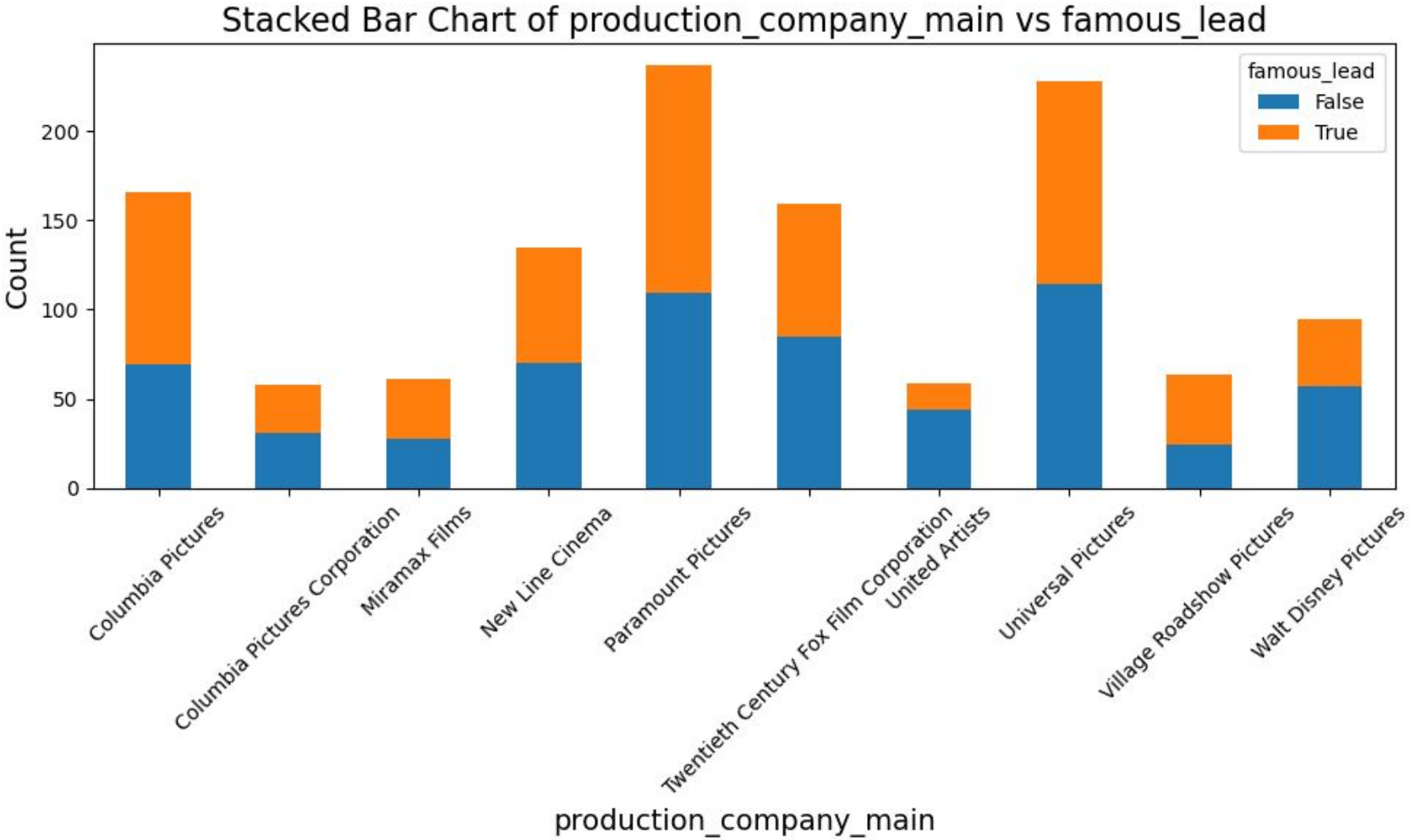


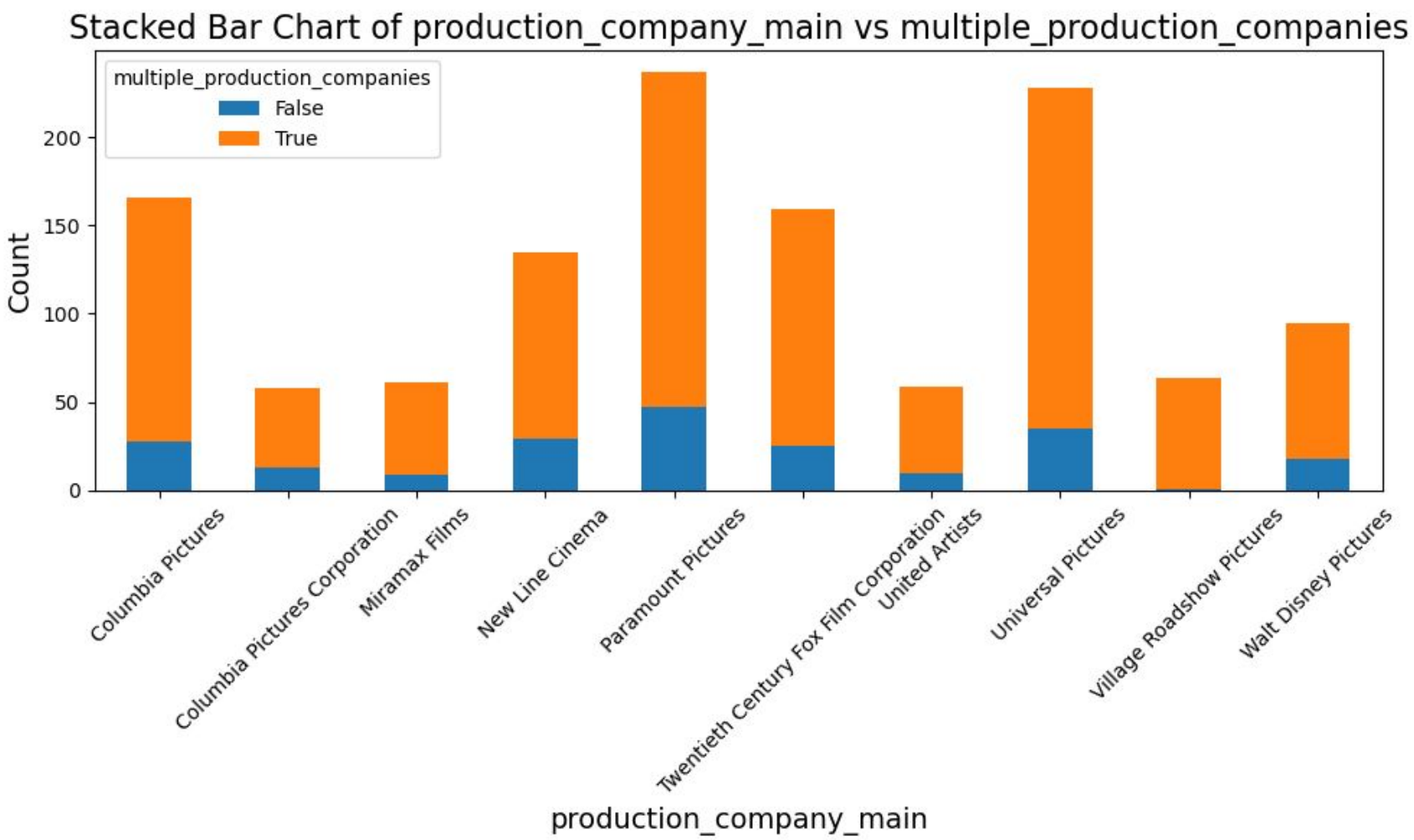


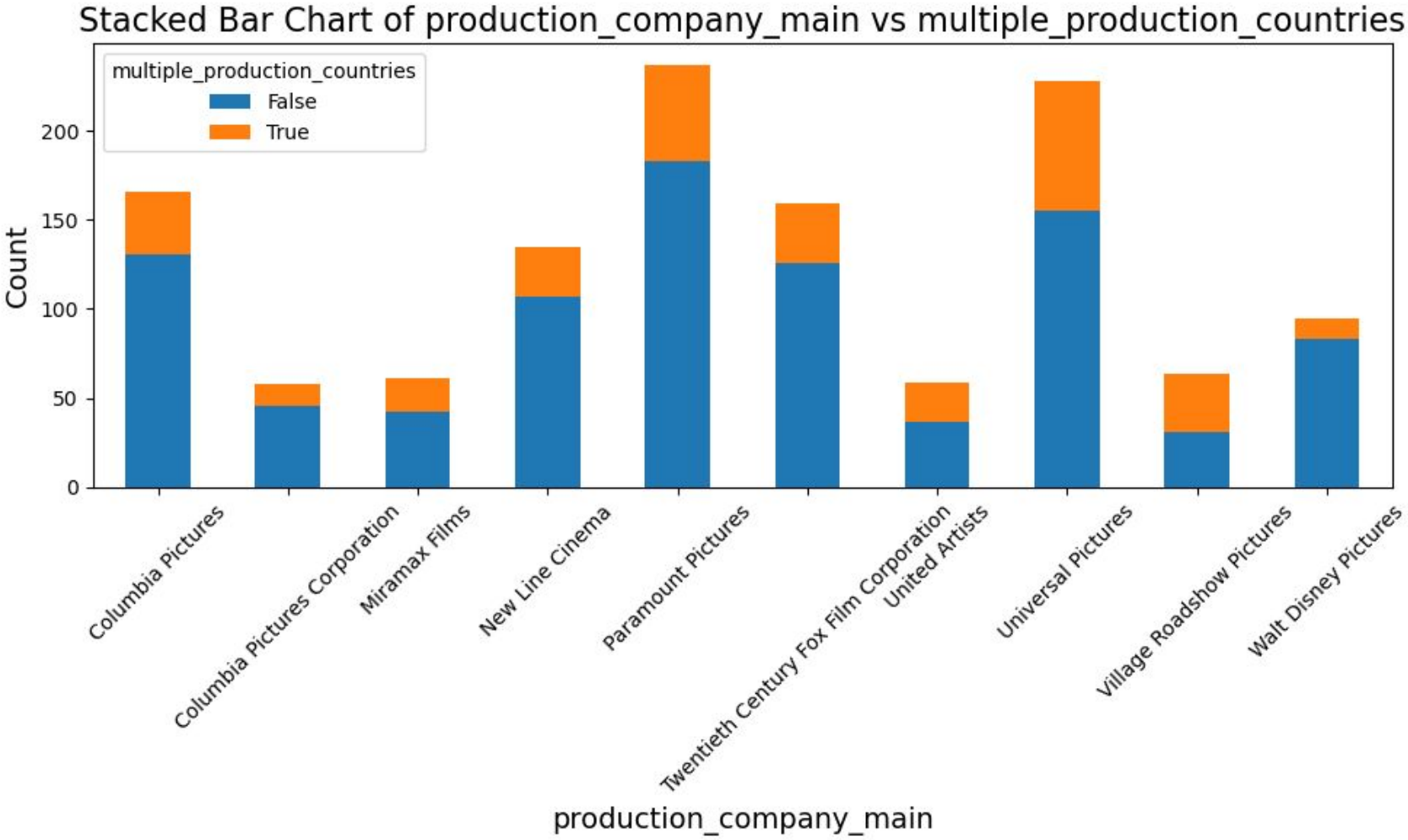


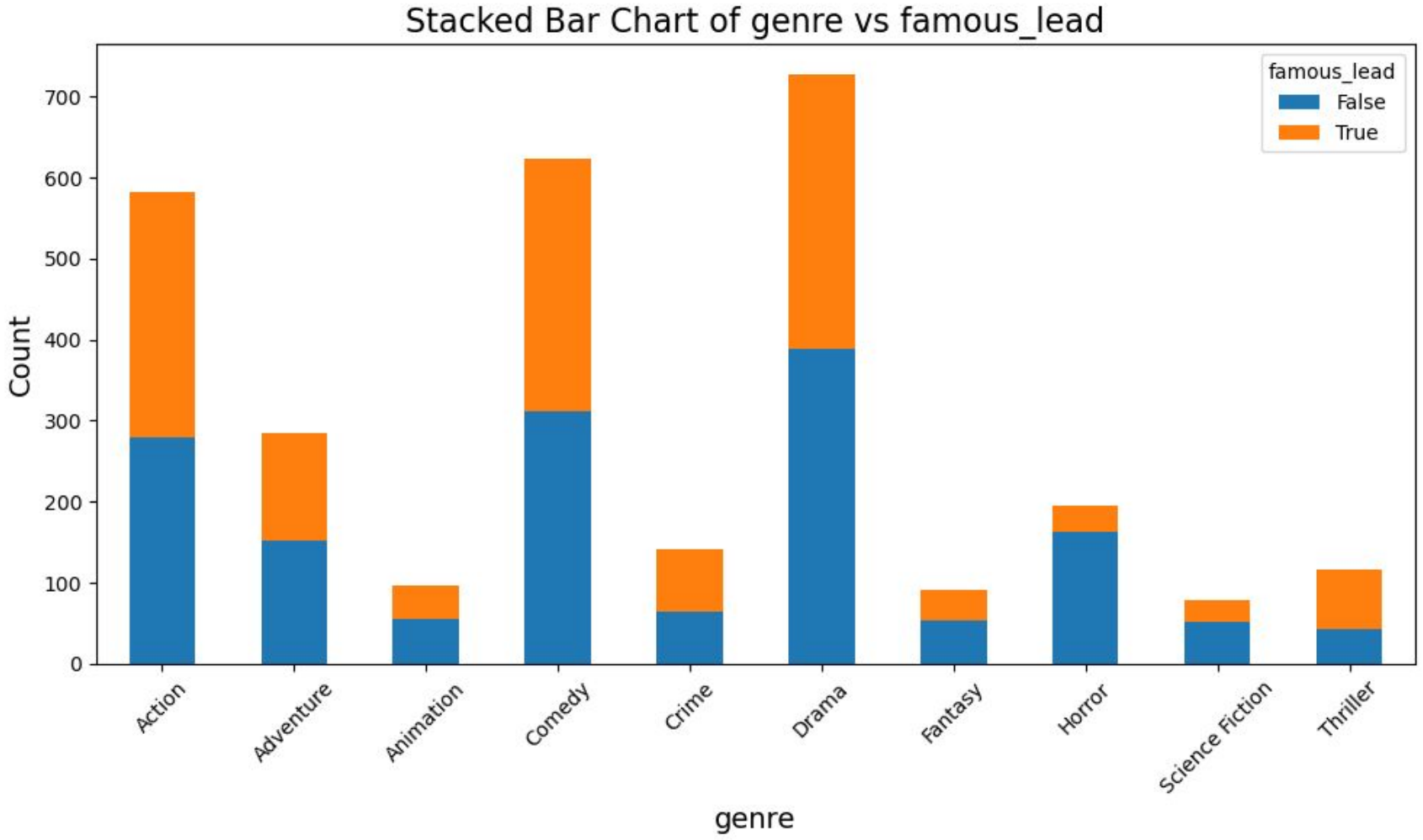
Bivariate Analysis (5)

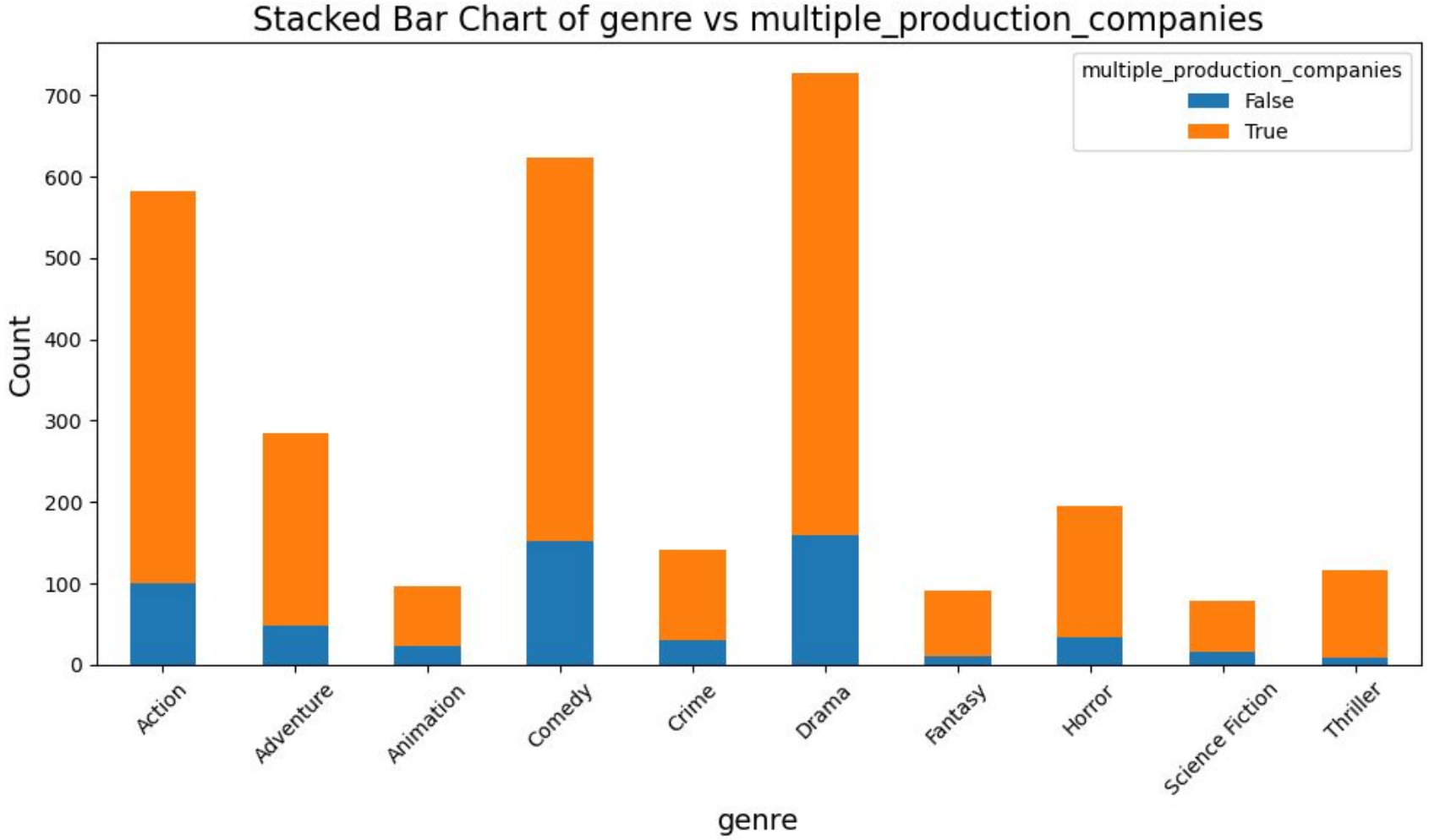
- The following slides contain stacked bar charts of categories against the binary/boolean categories, as others simply had too many features to properly visualize
  - Similar to the other visualizations; more insightful than anything

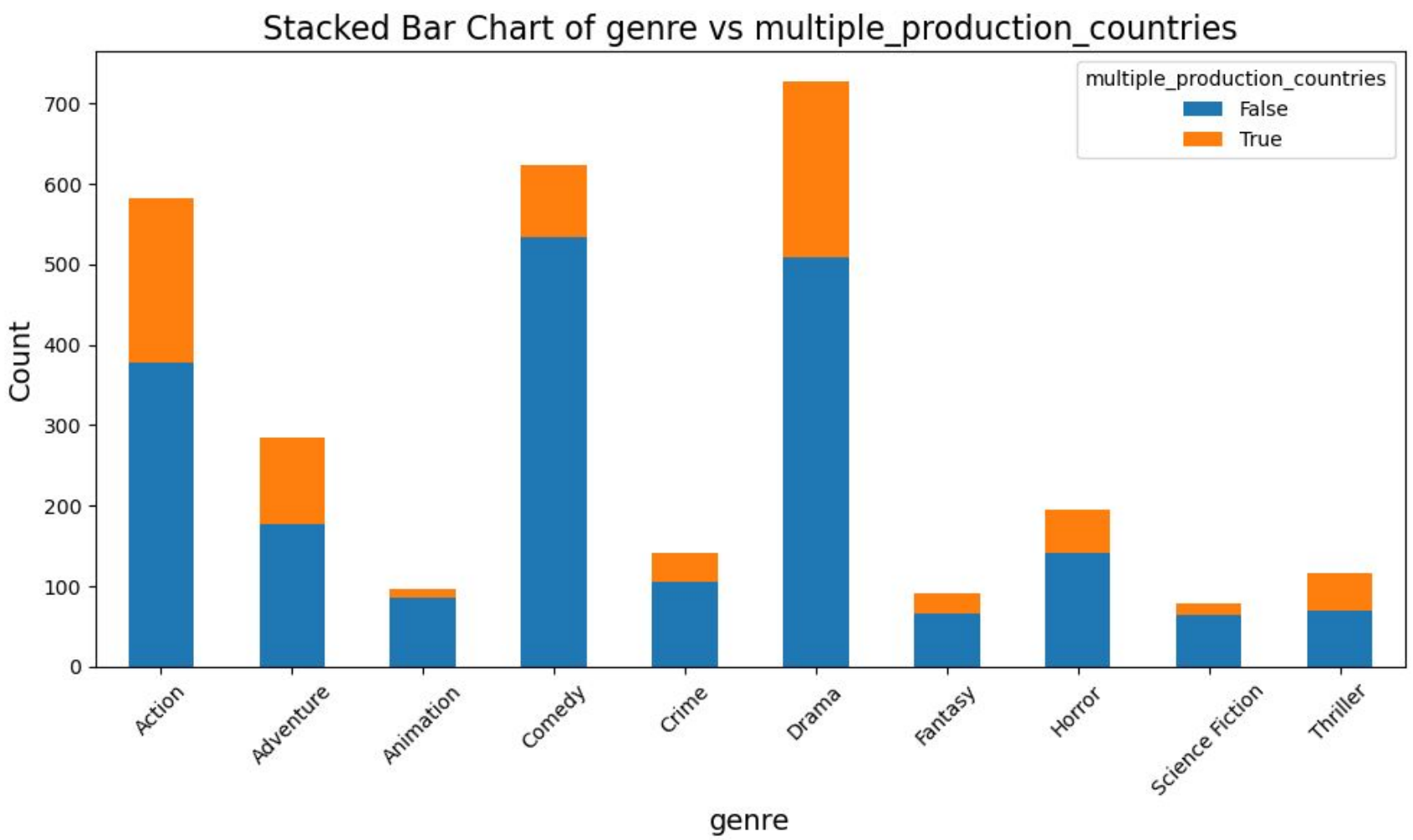

Stacked Bar Chart of production_company_main vs famous_lead
famous_lead
False
True
Count
0
50
100
150
200
Columbia Pictures
Columbia Pictures Corporation
Miramax Films
New Line Cinema
Paramount Pictures
Twentieth Century Fox Film Corporation
United Artists
Universal Pictures
Village Roadshow Pictures
Walt Disney Pictures
production_company_main

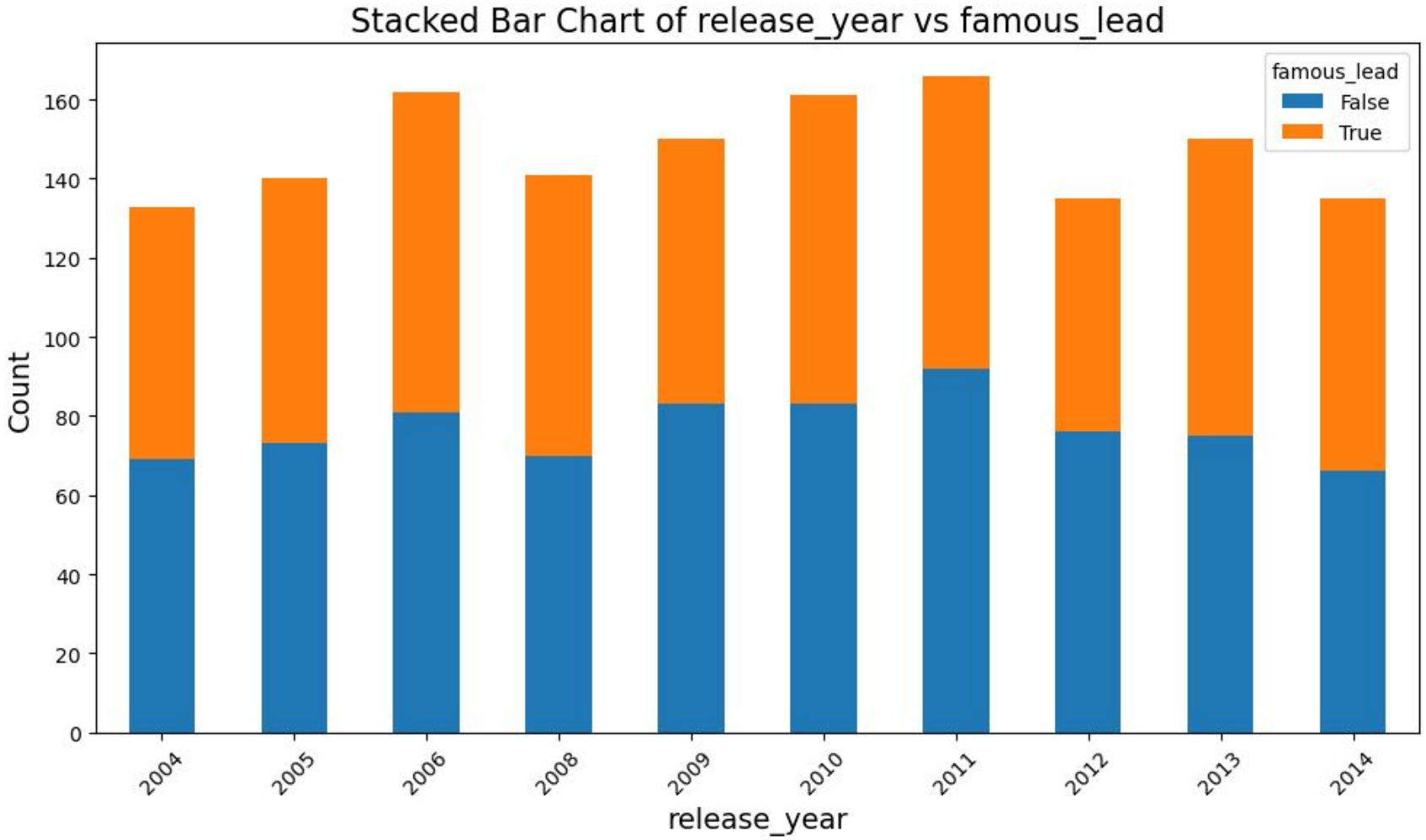

Stacked Bar Chart of production_company_main vs multiple_production_companies
multiple_production_companies
False
True
Count
0
50
100
150
200
Columbia Pictures
Columbia Pictures Corporation
Miramax Films
New Line Cinema
Paramount Pictures
Twentieth Century Fox Film Corporation
United Artists
Universal Pictures
Village Roadshow Pictures
Walt Disney Pictures
production_company_main

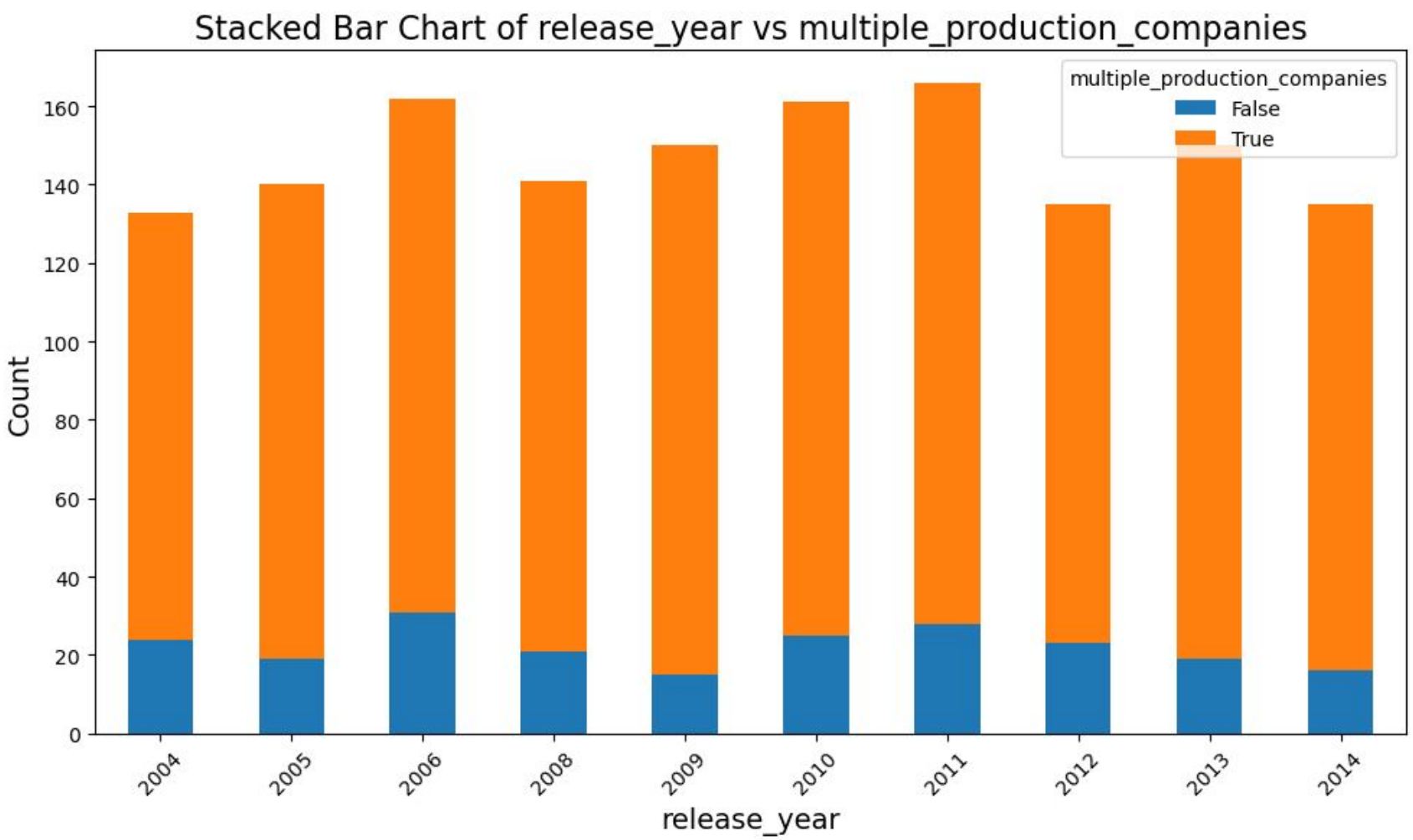

Stacked Bar Chart of production_company_main vs multiple_production_countries
multiple_production_countries
False
True
Count
0
50
100
150
200
Columbia Pictures
Columbia Pictures Corporation
Miramax Films
New Line Cinema
Paramount Pictures
Twentieth Century Fox Film Corporation
United Artists
Universal Pictures
Village Roadshow Pictures
Walt Disney Pictures
production_company_main

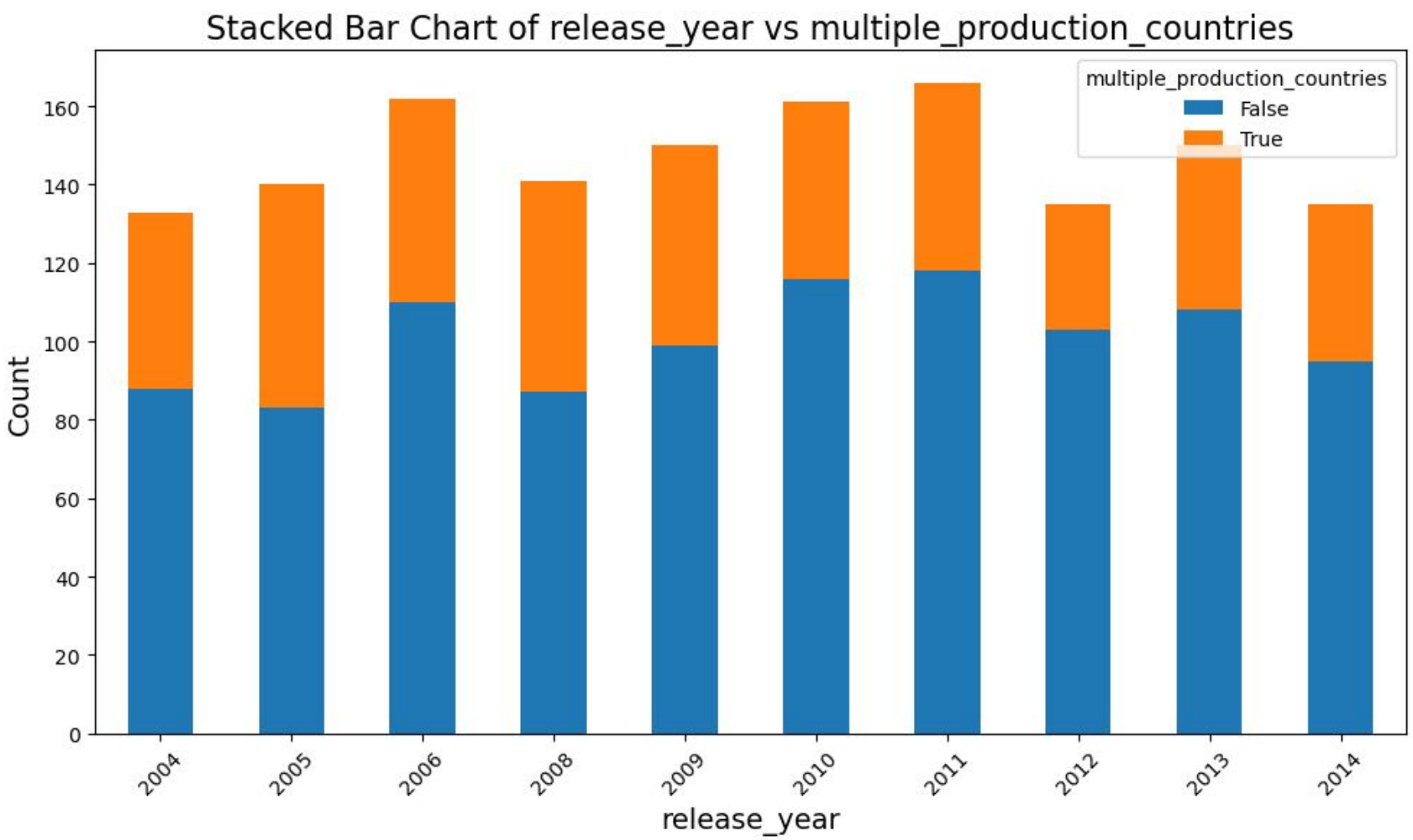

Stacked Bar Chart of genre vs famous_lead
famous_lead
False
True
Count
0
100
200
300
400
500
600
700
Action
Adventure
Animation
Comedy
Crime
Drama
Fantasy
Horror
Science Fiction
Thriller
genre

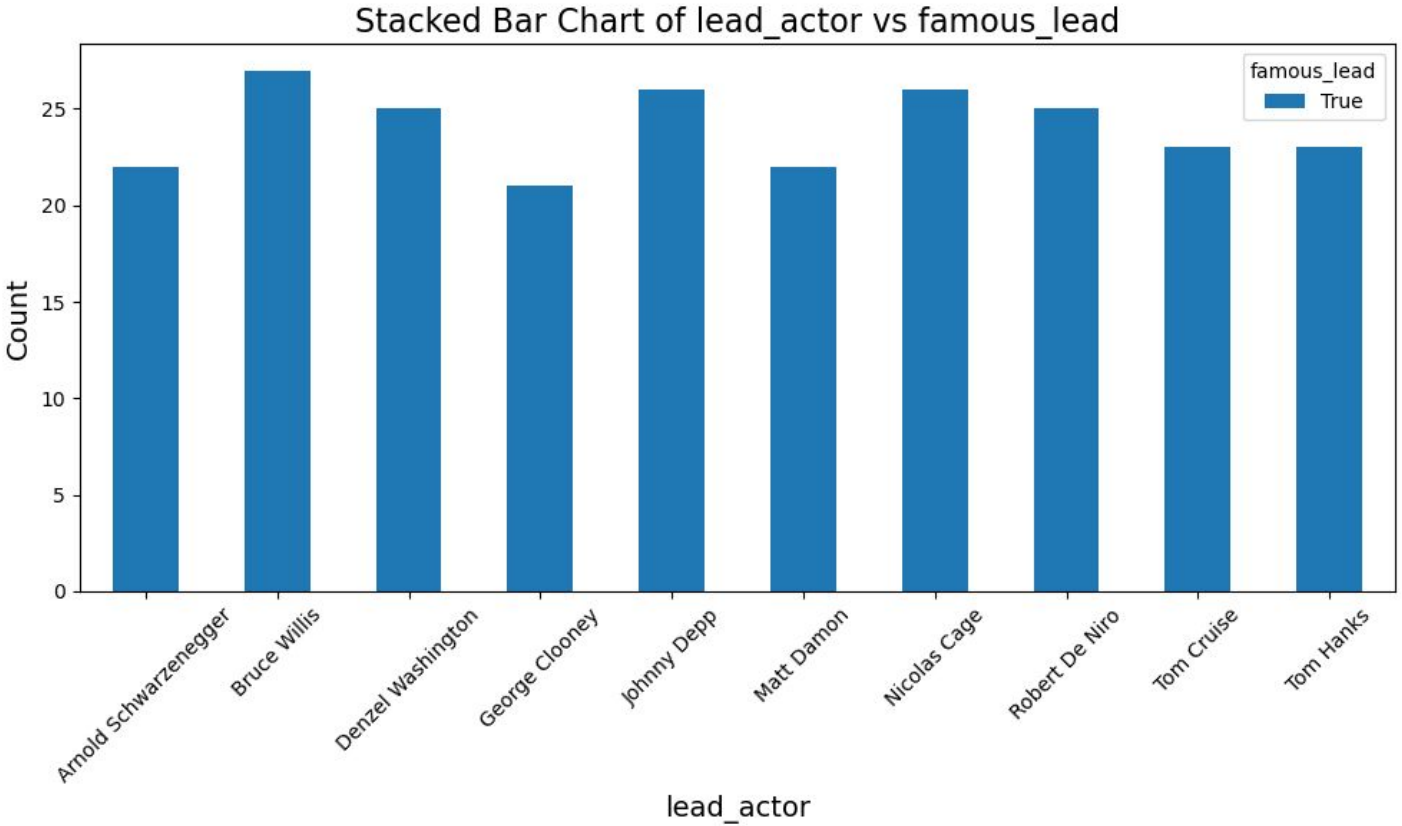

Stacked Bar Chart of genre vs multiple_production_companies
multiple_production_companies
False
True
Count
0
100
200
300
400
500
600
700
Action
Adventure
Animation
Comedy
Crime
Drama
Fantasy
Horror
Science Fiction
Thriller
genre


Stacked Bar Chart of genre vs multiple_production_countries
multiple_production_countries
False
True
Count
0
100
200
300
400
500
600
700
Action
Adventure
Animation
Comedy
Crime
Drama
Fantasy
Horror
Science Fiction
Thriller
genre

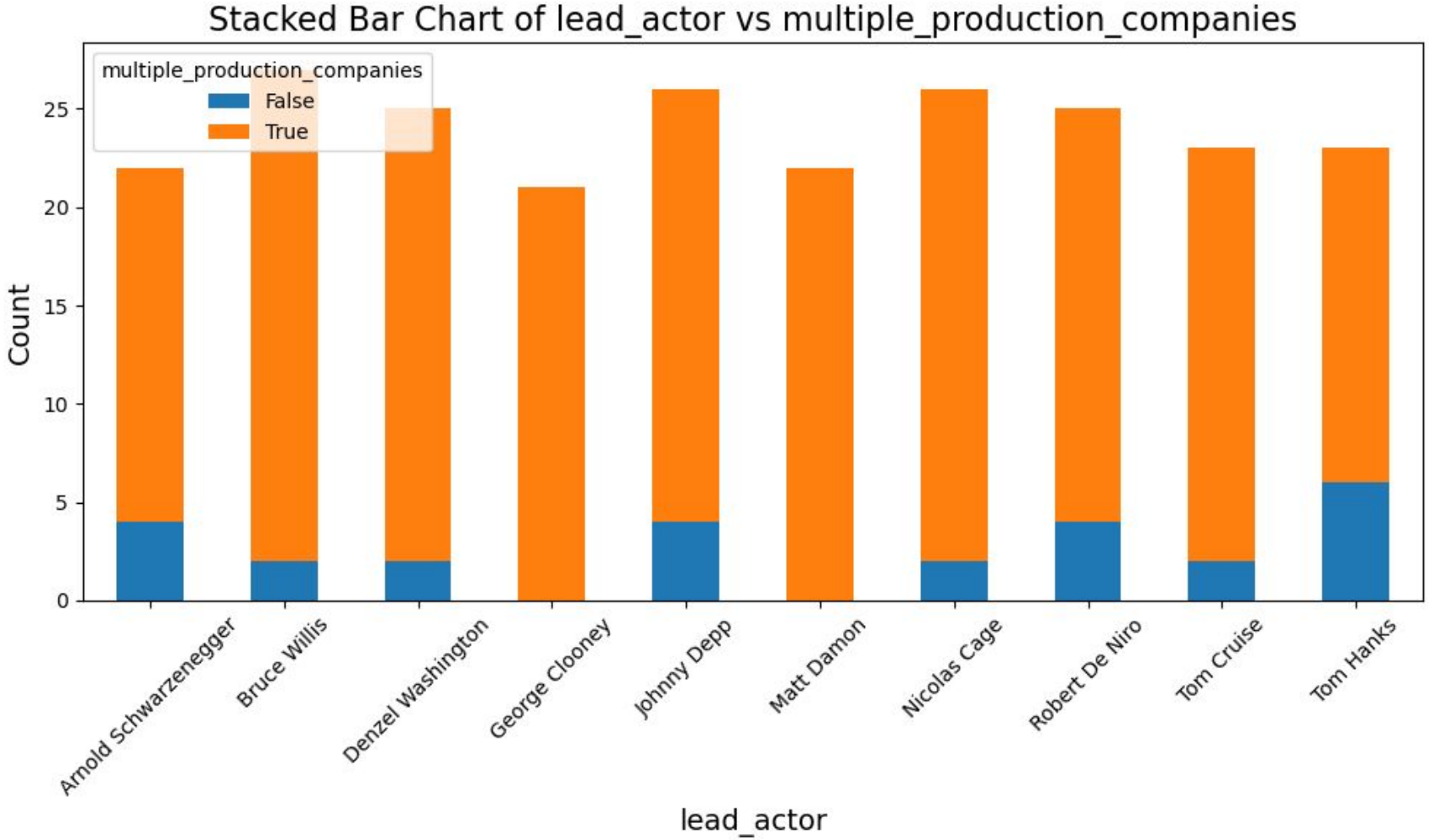

Stacked Bar Chart of release_year vs famous_lead
famous_lead
False
True
Count
0
20
40
60
80
100
120
140
160
2004
2005
2006
2008
2009
2010
2011
2012
2013
2014
release_year

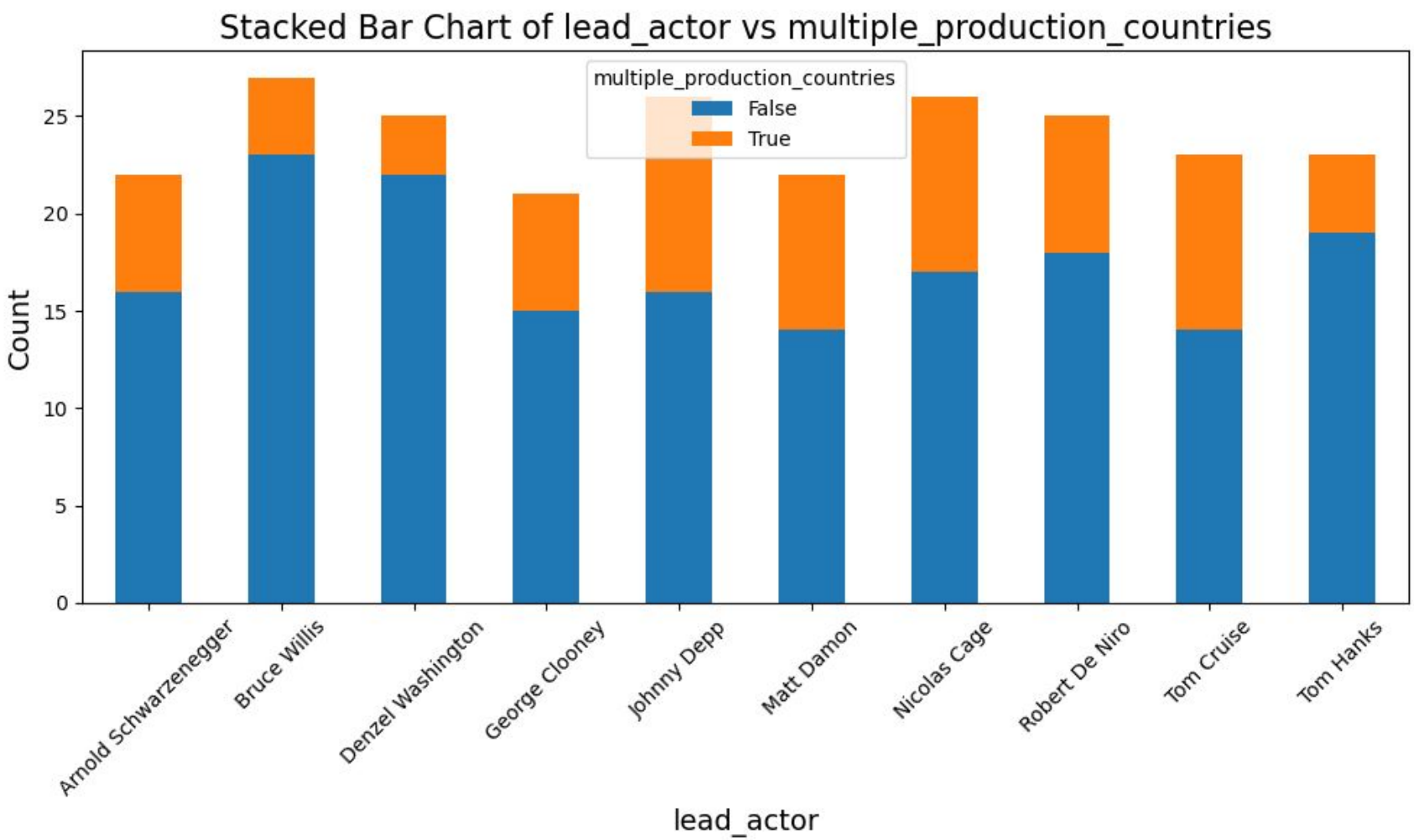

Stacked Bar Chart of release_year vs multiple_production_companies
multiple_production_companies
False
True
Count
0
20
40
60
80
100
120
140
160
2004
2005
2006
2008
2009
2010
2011
2012
2013
2014
release_year


Stacked Bar Chart of release_year vs multiple_production_countries
multiple_production_countries
False
True
Count
0
20
40
60
80
100
120
140
160
2004
2005
2006
2008
2009
2010
2011
2012
2013
2014
release_year



This one is funny but also good to confirm that the features are working as intended.

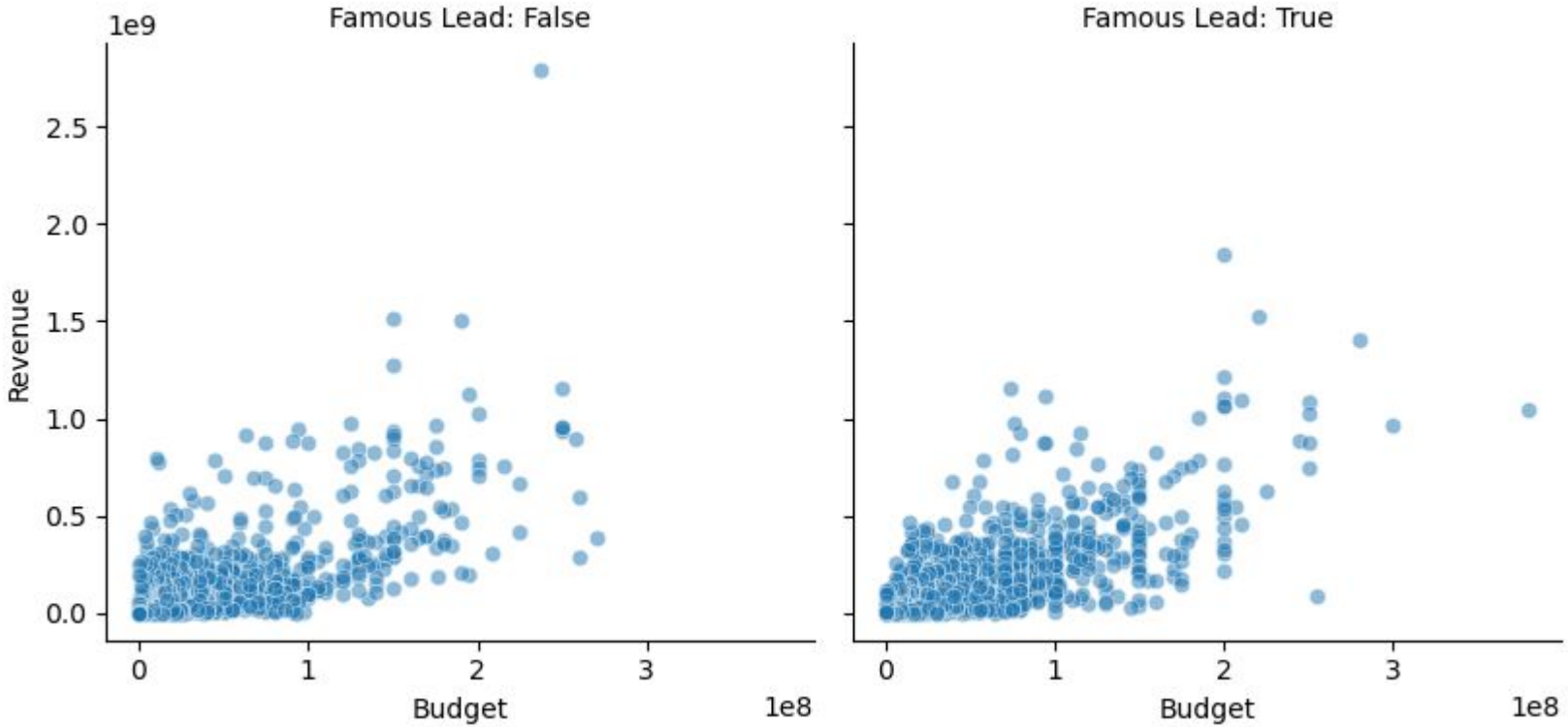

Stacked Bar Chart of lead_actor vs multiple_production_companies
multiple_production_companies
False
True
Count
0
5
10
15
20
25
Arnold Schwarzenegger
Bruce Willis
Denzel Washington
George Clooney
Johnny Depp
Matt Damon
Nicolas Cage
Robert De Niro
Tom Cruise
Tom Hanks
lead_actor

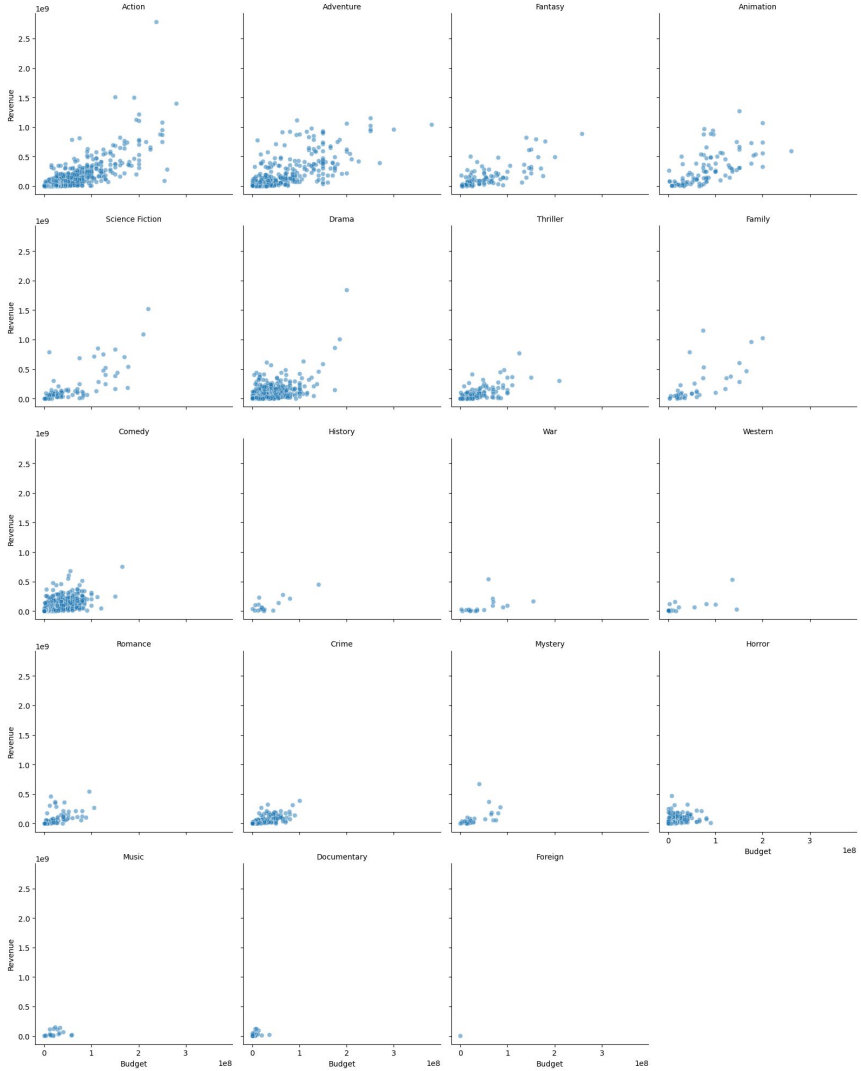

Stacked Bar Chart of lead_actor vs multiple_production_countries
multiple_production_countries
False
True
Count
0
5
10
15
20
25
Arnold Schwarzenegger
Bruce Willis
Denzel Washington
George Clooney
Johnny Depp
Matt Damon
Nicolas Cage
Robert De Niro
Tom Cruise
Tom Hanks
lead_actor

Multivariate Analysis

Also a bit of a difficult task to create meaningful visualizations because of the nature of the feature space

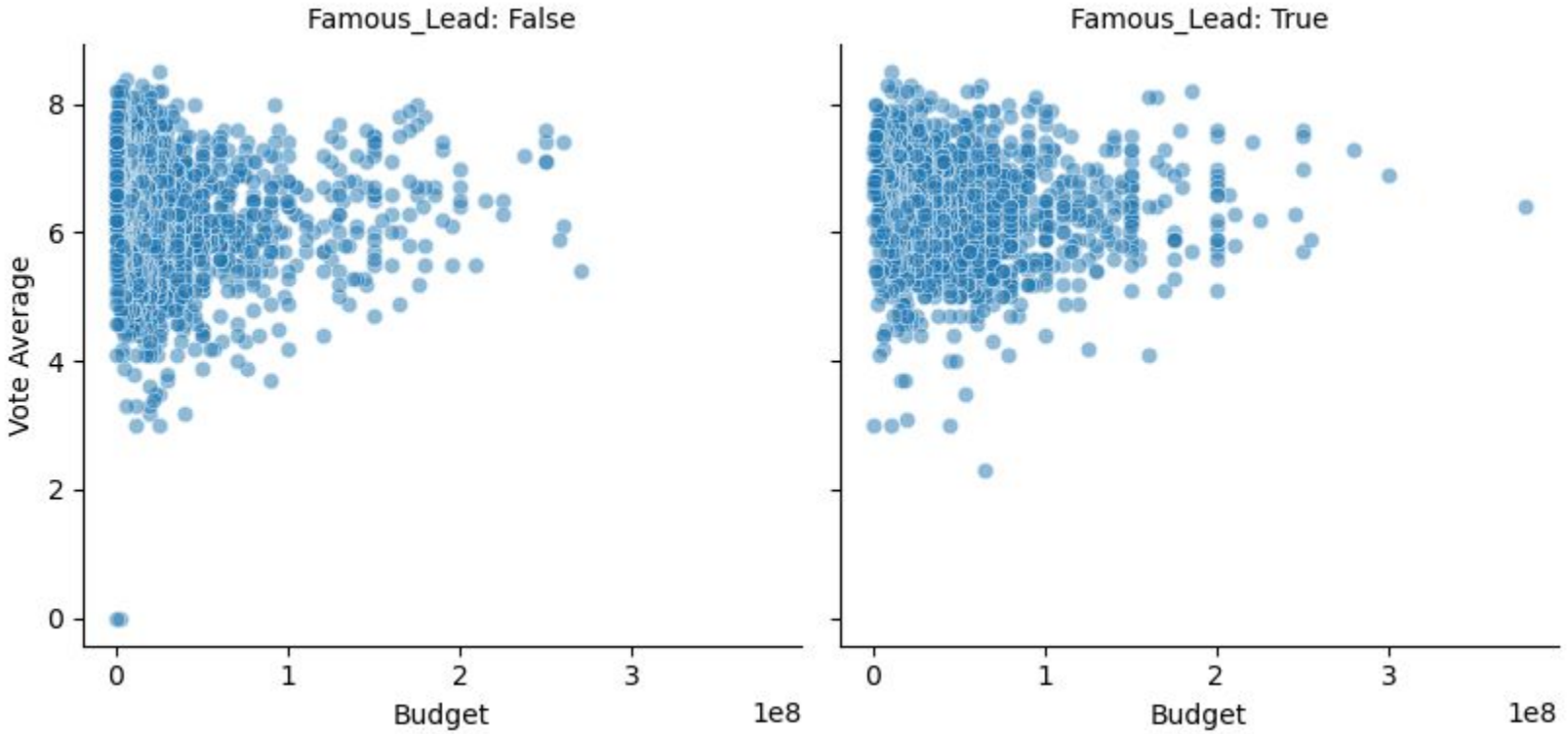
Faceted Scatterplot: Budget v Revenue v Famous_lead

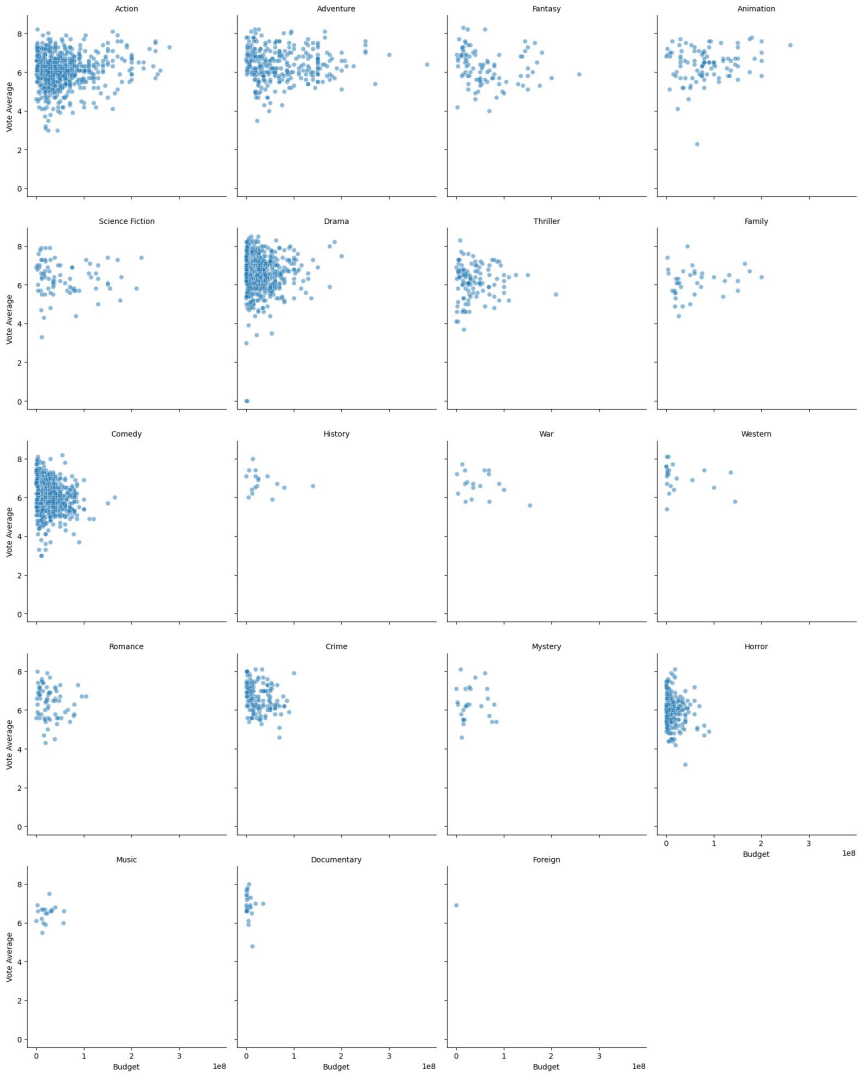




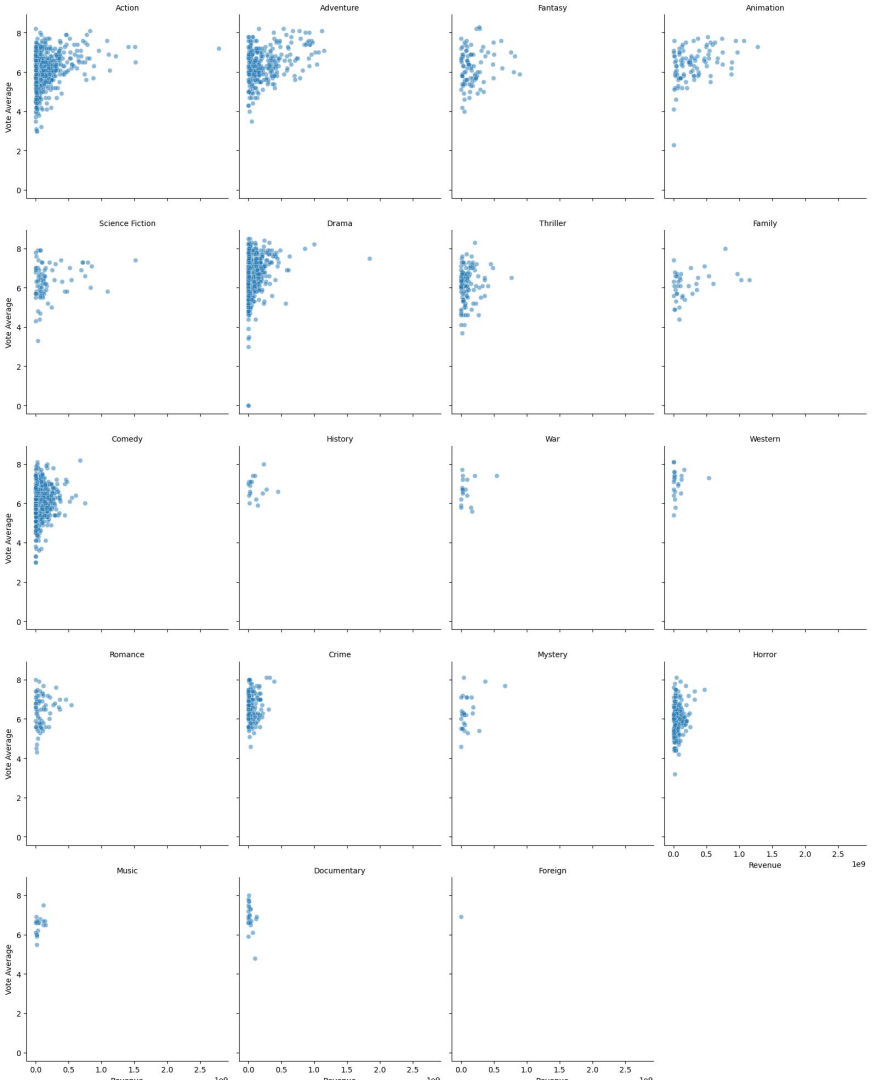
Faceted Scatterplot: Budget v Revenue v Genre

Faceted Scatterplot: Budget v vote_average v Famous_lead





Faceted Scatterplot: Budget v vote_average v Genre



Faceted Scatterplot: revenue v vote_average v Genre

Multivariate Diagnostics

Checking for outliers in multicollinearity:

- Using Mahalanobis Distance, only two movies were flagged as outliers; *Minions*, and *Interstellar*. Knowing both as movies, I understand why they were flagged, but they were kept.
- Variance inflation factor revealed that vote_count, as suspected, is the most likely to be highly correlated, however none of them were high enough to warrant any action.

	Variable	VIF
0	const	74.206056
1	budget	2.360788
2	popularity	2.282817
3	revenue	3.289296
4	runtime	1.308723
5	vote_average	1.503463
6	vote_count	3.896475
7	num_actors	1.267374
8	num_famous_actors	1.251970

Post-analysis Cleanup

- Changes I made
  - Renamed columns 'vote_average' to 'avg_user_score' and 'vote_count' to 'num_votes' for clarity purposes
  - Renamed ‘famous_lead’ to ‘has_famous_lead’ for clarity
    - Did this for the multiple_production_x features also
  - Added a ‘release_season’ feature based off of release_date for more detailed filtering opportunities
  - Added ‘decade_released’ for the same reason
  - Added inflation adjusted budget and revenue features for more accurate comparisons
  -

Final Dataset Summary

3174 rows x 26 columns

Numeric feature summary statistics:

	budget	id	popularity	release_date	revenue	runtime	avg_user_score	num_votes
count	3.174000e+03	3174.000000	3174.000000	3174	3.174000e+03	3174.000000	3174.000000	3174.000000
mean	4.129071e+07	44883.733144	29.475862	2002-04-01 19:41:51.153119104	1.232553e+08	110.897290	6.312161	993.340895
min	7.000000e+03	5.000000	0.037073	1916-09-04 00:00:00	7.000000e+00	41.000000	0.000000	0.000000
25%	1.100000e+07	4871.250000	10.926848	1998-08-20 06:00:00	1.786998e+07	97.000000	5.800000	190.000000
50%	2.600000e+07	11357.500000	20.801659	2005-08-04 12:00:00	5.722771e+07	107.000000	6.300000	485.000000
75%	5.500000e+07	44944.500000	37.764689	2010-11-29 06:00:00	1.491741e+08	121.000000	6.900000	1168.750000
max	3.800000e+08	417859.000000	875.581305	2016-09-09 00:00:00	2.787965e+09	338.000000	8.500000	13752.000000
std	4.450746e+07	75070.163522	36.314834	NaN	1.872728e+08	20.966473	0.869925	1421.112279

	num_actors	num_famous_actors	release_year	decade_released	2024_adjusted_budget	2024_adjusted_revenue
count	3174.000000	3174.000000	3174.000000	3174.000000	3143.0	3143.0
mean	26.281664	2.204159	2001.712350	1997.375551	69455054.035317	219199349.603882
std	21.558386	2.038837	13.243916	13.659660	67250620.445985	326163763.411103
min	1.000000	0.000000	1916.000000	1910.000000	11158.0	13.0
25%	15.000000	1.000000	1998.000000	1990.000000	21896050.5	34841355.5
50%	19.000000	2.000000	2005.000000	2000.000000	47768488.0	105590220.0
75%	30.000000	3.000000	2010.000000	2010.000000	96390594.0	261755021.0
max	224.000000	15.000000	2016.000000	2010.000000	529982157.0	4076060708.0

Final Dataset Summary (2)

Measure summaries:

	count	unique	top	freq
original_language	3174	25	en	3053
original_title	3174	3174	Avatar	1
release_date	3174	NaN	NaN	NaN
title	3174	3173	The Host	2
director	3174	1410	Steven Spielberg	27
genre	3174	19	Drama	728
lead_actor	3174	1266	Bruce Willis	27
has_famous_lead	3174	2	False	1703
production_company_main	3174	748	Paramount Pictures	237
has_multiple_production_companies	3174	2	True	2540
production_country_main	3174	45	United States of America	2222
has_multiple_production_countries	3174	2	False	2296
season_releaed	3174	4	Fall	913

Final Data Dictionary

Variable Name	Data Type	Description
id	Integer	Unique movie identifier
budget	Integer	Budget of the movie (USD)
2024_adjusted_budget	Integer	Budget of the movie adjusted for 2024 CPI
revenue	Integer	Revenue the movie generated (in USD)
2024_adjusted_revenue	Integer	Revenue generated adjusted for 2024 CPI
original_language	String	ISO-639-1 code for the original language
original_title	String	Title of the movie in its original language
title	String	Title of the movie
Release_date	datetime	Date when movie was first released

Final Data Dictionary (2)

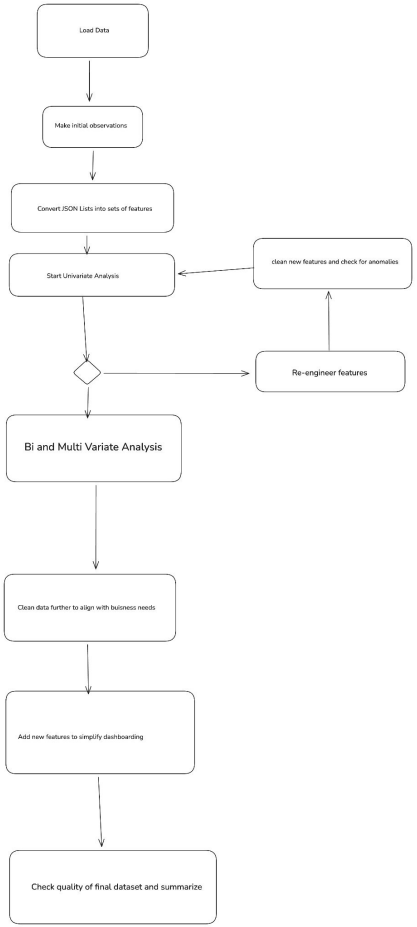
Variable Name	Data Type	Description
runtime	Integer	Length of the movie (minutes
avg_user_score	float	Average user rating of the movie
num_votes	Integer	Number of votes by users on the movie
popularity	float	Popularity score assigned by tMDB
genre	String	Main genre of the movie
director	String	Director of the movie
Lead_actor	String	Lead actor of movie
num_actors	Integer	Number of actors in the movie
num_famous_actors	integer	Number of actors who have appeared in >= 20 movies in the database

Final Data Dictionary (3)

Variable Name	Data Type	Description
has_famous_lead	boolean	Is the lead actor also famous
production_company_main	String	Main company that produced the movie
has_multiple_production_companies	boolean	If multiple companies worked on the movie
production_country_main	String	Main country the movie was produced in
has_multiple_production_countries	boolean	If it was produced in multiple countries
release_year	Integer	What year the movie came out in
season_releaed	String	What season the movie was released during
decade_released	Integer	What decade the movie released in

Workflow Summary

Basic flowchart I made:



Code Repository

Notebook used for process as well as files data came from can be found at

https://github.com/japhelan/tMDB-analysis

Next Steps

Most features are engineered with dashboard implementation in mind.

- If used for modeling, select only one of the features within release_date, release_year, and release_decade, as they have high correlation
  - Same applies to inflation adjusted statistics and their non-adjusted counterparts
- Almost all the meaningful visualizations were put into the presentation, but more can be found on in the notebook on GitHub or in the dashboard