بخش اول شناخت داده:

کدی که ارائه شده، تابعی به نام calculate_statisticsرا تعریف می‌کند که برای محاسبه‌ی آمار توصیفی ویژگی‌های عددی در یک DataFrameاستفاده می‌شود. در ادامه توضیحات کامل درباره‌ی مراحل این کد و تحلیل خروجی ارائه می‌گردد:

مراحل انجام کار:

1. **تعریف تابع :calculate_statistics**
   - این تابع به دو ورودی نیاز دارد data_frame :که داده‌های ورودی را شامل می‌شود و Numeric_Featuresکه لیستی از ویژگی‌های عددی برای تحلیل است.
   - در هر تکرار حلقه، برای هر ویژگی عددی آماره‌هایی از جمله کمینه(Min) ، بیشینه(Max) ، میانگین(Mean) ، میانه (Median)، مد(Mode) ، و محدوده‌ی مقادیر (Range) محاسبه می‌شود.
2. **محاسبه‌ی آماره‌ها:**
   - **نوع داده :(dtype)** نوع داده‌ی هر ستون مشخص می‌شود تا نوع مقادیر ذخیره شده در آن را ببینیم.
   - **کمینه و بیشینه :(Min, Max)** کم‌ترین و بیش‌ترین مقدار در هر ستون مشخص می‌شود.
   - **میانگین :(Mean)** میانگین مقادیر هر ستون، نشان‌دهنده‌ی مقدار متوسط است.
   - **میانه :(Median)** میانه، مقداری است که داده‌ها را به دو نیمه تقسیم می‌کند.
   - **مد :(Mode)** مقداری که بیشترین تکرار را دارد.
   - **محدوده :(Range)** نشان‌دهنده‌ی بازه‌ی بین کمینه و بیشینه‌ی هر ستون است.
3. **تبدیل به:DataFrame**
   - نتایج به دست آمده در قالب یک DataFrame ذخیره می‌شوند تا قابل نمایش و تفسیر باشند.

روش انتخاب شده و دلیل آن:

استفاده از آماره‌های توصیفی (مانند میانگین، میانه و مد) به ما کمک می‌کند تا درک بهتری از توزیع داده‌ها داشته باشیم. در اینجا، انتخاب این روش به این دلیل بوده است که:

- **شناسایی داده‌های پرت :(Outliers)** مقادیر کمینه و بیشینه می‌توانند به شناسایی داده‌های غیرعادی کمک کنند.
- **توزیع داده‌ها** :میانگین و میانه می‌توانند نشان دهند که داده‌ها چگونه در ستون‌ها توزیع شده‌اند و آیا انحرافی در داده‌ها وجود دارد یا خیر.
- **تحلیل کلی** :این آمارها به عنوان قدم اول برای تحلیل داده‌ها و شناسایی مشکلات احتمالی مانند داده‌های گم‌شده یا توزیع نامتقارن استفاده می‌شوند.

6 rows × 8 columns

	Feature	Type	Min	Max	Mean	Median	Mode	Range
0	vote_average	float64	0.0	1.000000e+01	2.597083e+00	0.0	0.0	0.0 - 10.0
1	vote_count	int64	0.0	1.946300e+04	4.039538e+01	0.0	0.0	0 - 19463
2	runtime	int64	0.0	3.720000e+03	2.090141e+01	7.0	0.0	0 - 3720
3	revenue	int64	0.0	1.450027e+09	1.863464e+06	0.0	0.0	0 - 1450026933
4	budget	int64	0.0	2.600000e+08	5.759513e+05	0.0	0.0	0 - 260000000
5	popularity	float64	0.0	1.008942e+03	1.886233e+00	0.6	0.6	0.0 - 1008.942

تحلیل خروجی:

در خروجی نشان داده شده:

- برای ستون vote_average:
  - میانگین امتیاز پایین (حدود 2.6) و میانه صفر نشان می‌دهد که بسیاری از فیلم‌ها امتیاز پایین یا حتی صفر دریافت کرده‌اند.
- برای ستون vote_count:
  - میانگین تعداد رأی‌ها حدود 40 است، اما میانه و مد صفر نشان می‌دهد که تعداد زیادی از فیلم‌ها هیچ رأیی دریافت نکرده‌اند.
- برای ستون runtime:
  - وجود مقادیر کمینه صفر و بیشینه 3720 نشان‌دهنده‌ی داده‌های نامعتبر یا اشتباه است (مانند مدت زمان بسیار زیاد برای یک فیلم).
- برای ستون‌های revenue و budget:
  - وجود مقادیر صفر در مد و میانه نشان‌دهنده‌ی تعداد زیاد فیلم‌هایی است که داده‌های مالی ثبت نشده یا صفر دارند.
- برای ستون popularity:
  - میانگین پایین و بیشینه بالای 1000 نشان می‌دهد که تعداد کمی از فیلم‌ها محبوبیت بسیار بالایی داشته‌اند، در حالی که اکثر فیلم‌ها محبوبیت پایینی داشته‌اند.

نتیجه‌گیری:

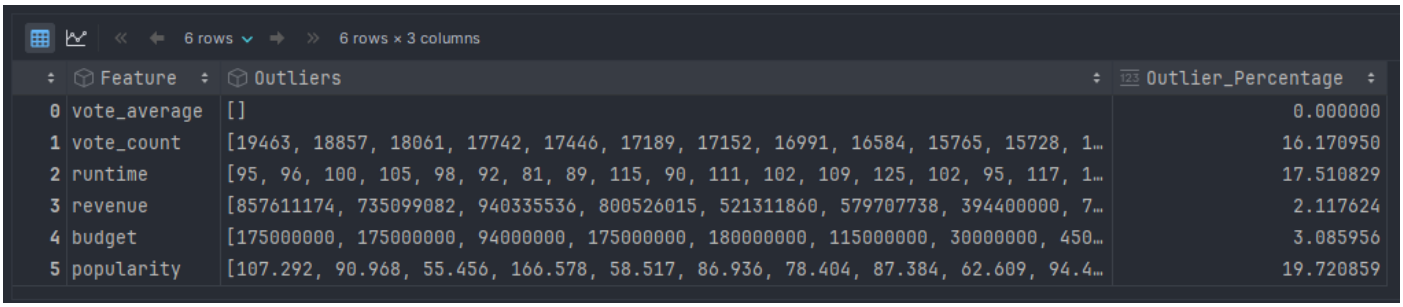
این تحلیل اولیه نشان می‌دهد که برخی از ستون‌ها دارای داده‌های گم‌شده یا اشتباه هستند و نیاز به پیش‌پردازش و پاکسازی داده‌ها دارند.

کد برای شناسایی داده‌های پرت در ویژگی‌های عددی Numeric_Features یک DataFrame به نام outliers_summary تولید می‌کند که حاوی اطلاعات زیر است:

مراحل انجام شده در کد:

1. **محاسبه چارک‌ها و IQR:**
   - برای هر ویژگی، چارک اول (Q1) و چارک سوم (Q3) محاسبه می‌شوند.
   - **IQR (Interquartile Range)** به عنوان تفاوت بین Q3 و Q1 محاسبه می‌شود.
2. **تشخیص داده‌های پرت:**
   - مقادیری که کمتر از $Q1 - 1.5 \times \text{IQR}$ یا بیشتر از $Q3 + 1.5 \times \text{IQR}$ هستند به عنوان داده‌های پرت شناخته می‌شوند.

3. **محاسبه درصد داده‌های پرت:**
   - درصد داده‌های پرت با تقسیم تعداد داده‌های پرت بر تعداد کل داده‌ها در هر ویژگی محاسبه می‌شود.
4. **ایجاد DataFrame نهایی:**
   - ویژگی‌های شناسایی شده به همراه مقادیر داده‌های پرت و درصد آن‌ها در یک DataFrame نهایی ذخیره می‌شوند.



6 rows × 3 columns

	Feature	Outliers	Outlier_Percentage
0	vote_average	[]	0.000000
1	vote_count	[19463, 18857, 18061, 17742, 17446, 17189, 17152, 16991, 16584, 15765, 15728, 1…	16.170950
2	runtime	[95, 96, 100, 105, 98, 92, 81, 89, 115, 90, 111, 102, 109, 125, 102, 95, 117, 1…	17.510829
3	revenue	[857611174, 735099082, 940335536, 800526015, 521311860, 579707738, 394400000, 7…	2.117624
4	budget	[175000000, 175000000, 94000000, 175000000, 180000000, 115000000, 30000000, 450…	3.085956
5	popularity	[107.292, 90.968, 55.456, 166.578, 58.517, 86.936, 78.404, 87.384, 62.609, 94.4…	19.720859

تحلیل خروجی:

- **Feature**: نام ویژگی عددی.
- **Outliers**: مقادیر شناسایی شده به عنوان داده‌های پرت در هر ویژگی.
- **Outlier_Percentage**: درصد داده‌های پرت نسبت به کل داده‌ها برای هر ویژگی.

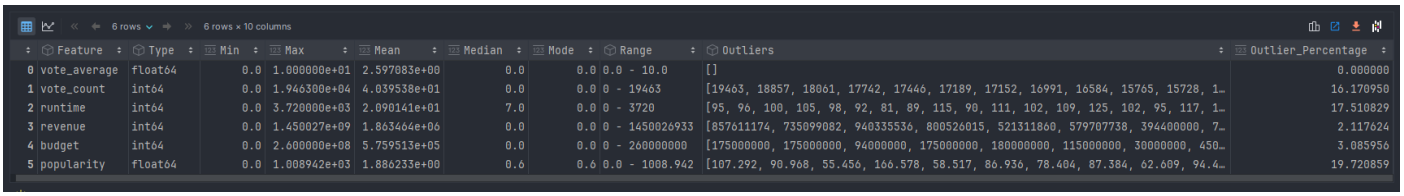
بررسی خروجی:

1. **vote_average**:
   - بدون داده پرت.(0%)
2. **vote_count**:
   - حدود 16.17٪ داده‌ها به عنوان داده پرت شناسایی شده‌اند. این نشان می‌دهد که تعداد زیادی از فیلم‌ها تعداد آرای بسیار بالا یا پایین دارند.
3. **runtime**:
   - حدود 17.51٪ داده‌ها به عنوان داده پرت شناسایی شده‌اند. مقادیر پرت در این ویژگی می‌تواند نشان‌دهنده مقادیر غیرعادی مانند زمان‌های بسیار کوتاه یا بلند باشد.
4. **revenue**:
   - حدود 2.12٪ داده‌ها به عنوان داده پرت شناسایی شده‌اند. این مقادیر می‌تواند مربوط به فیلم‌هایی باشد که درآمد بسیار بالایی داشته‌اند.
5. **budget**:
   - حدود 3.09٪ داده‌ها به عنوان داده پرت شناسایی شده‌اند، که می‌تواند به دلیل بودجه‌های غیرعادی بالا باشد.
6. **popularity**:
   - حدود 19.72٪ داده‌ها به عنوان داده پرت شناسایی شده‌اند، که نشان می‌دهد برخی از فیلم‌ها دارای محبوبیت بسیار بالایی نسبت به بقیه هستند.

نتیجه‌گیری:

تحلیل داده‌های پرت می‌تواند به ما در شناسایی مقادیر غیرعادی کمک کند که ممکن است نیاز به پاکسازی یا بررسی بیشتری داشته باشند. این داده‌های پرت می‌توانند تأثیر زیادی بر نتایج تحلیل و مدل‌های یادگیری ماشین داشته باشند و باید به دقت مدیریت شوند.

Final Tabel

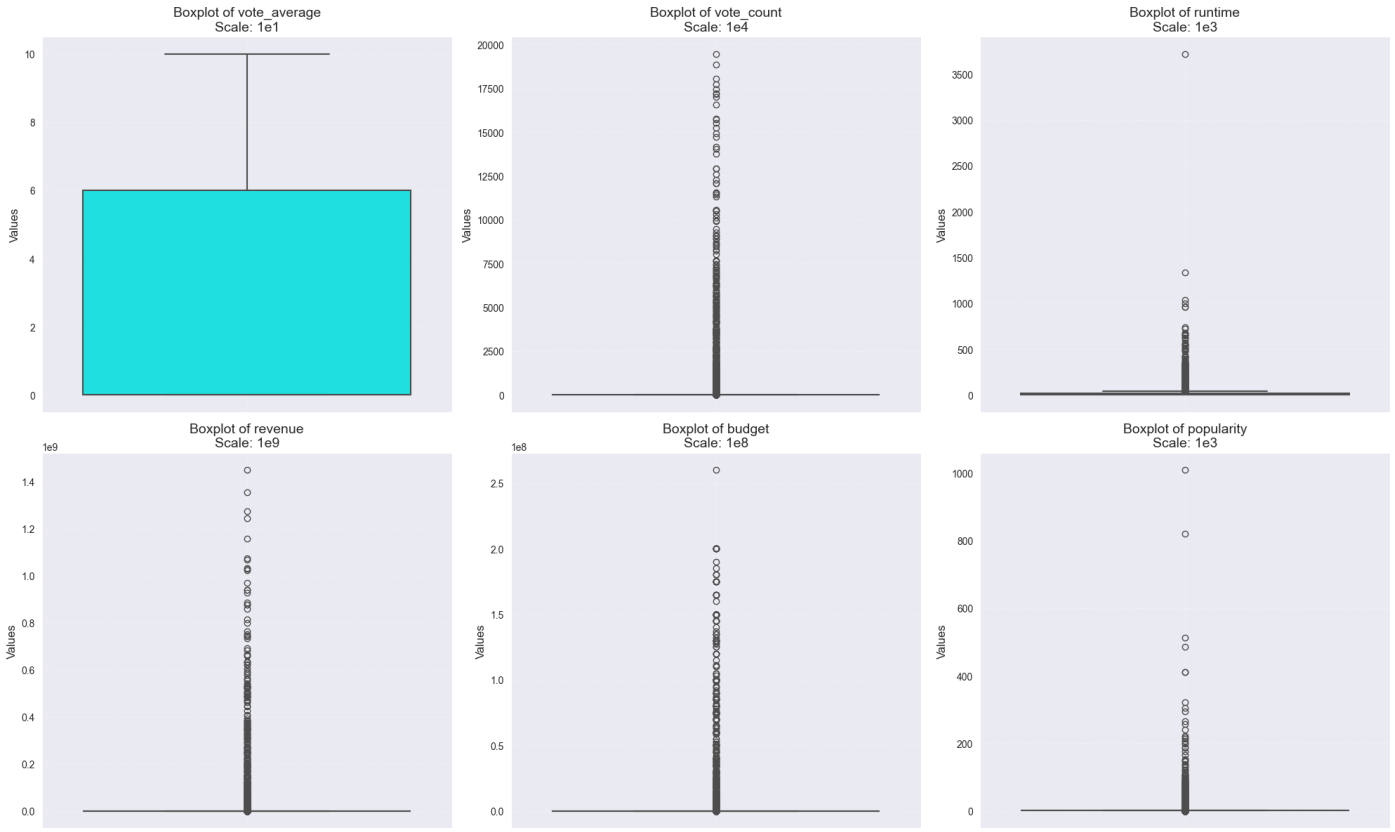


	Feature	Type	Min	Max	Mean	Median	Mode	Range	Outliers	Outlier_Percentage
0	vote_average	float64	0.0	1.000000e+01	2.597083e+00	0.0	0.0	0.0 - 10.0	[]	0.000000
1	vote_count	int64	0.0	1.946300e+04	4.039538e+01	0.0	0.0	0 - 19463	[19463, 18857, 18061, 17742, 17446, 17189, 17152, 16991, 16584, 15765, 15728, 1…	16.170950
2	runtime	int64	0.0	3.720000e+03	2.090141e+01	7.0	0.0	0 - 3720	[95, 96, 100, 105, 98, 92, 81, 89, 115, 90, 111, 102, 109, 125, 102, 95, 117, 1…	17.510829
3	revenue	int64	0.0	1.450027e+09	1.863464e+06	0.0	0.0	0 - 1450026933	[857611174, 735099082, 940335536, 800526015, 521311860, 579707738, 394400000, 7…	2.117624
4	budget	int64	0.0	2.600000e+08	5.759513e+05	0.0	0.0	0 - 260000000	[175000000, 175000000, 94000000, 175000000, 180000000, 115000000, 30000000, 450…	3.085956
5	popularity	float64	0.0	1.008942e+03	1.886233e+00	0.6	0.6	0.0 - 1008.942	[107.292, 90.968, 55.456, 166.578, 58.517, 86.936, 78.404, 87.384, 62.609, 94.4…	19.720859

نمودار Boxplot برای داده های عددی

توضیح مراحل کد:

1. **تعریف تابع :plot_boxplots**
   - تابع ورودی‌های مختلفی از جمله DataFrame ، لیست ویژگی‌ها (ستون‌های عددی)، و یک لیست از ستون‌هایی که می‌خواهیم مقادیر صفر آن‌ها را در نظر بگیریم، می‌پذیرد.
   - پارامتر after_preprocessing برای فیلتر کردن داده‌های پرت (مخصوصاً در ویژگی runtime) استفاده می‌شود.
2. **محاسبه مقیاس علمی:**
   - حداکثر مقدار مطلق در هر ستون محاسبه شده و برای نمایش مقیاس علمی (scientific notation) استفاده می‌شود.
3. **ایجاد نمودار جعبه‌ای:**
   - برای هر ویژگی، نمودار جعبه‌ای رسم می‌شود.
   - در صورتی که ستون جزو exclude_zero_columns باشد، مقادیر صفر به عنوان داده پرت نمایش داده می‌شوند.
   - اگر پارامتر after_preprocessing فعال باشد و ستون مورد نظر runtime باشد، داده‌ها فیلتر شده و داده‌های پرت بر اساس محدوده خاصی نمایش داده می‌شوند.
4. **نمایش خروجی:**
   - نمودارهای جعبه‌ای برای هر ویژگی نمایش داده می‌شوند که در آن‌ها مقیاس علمی و فیلترها اعمال شده‌اند.



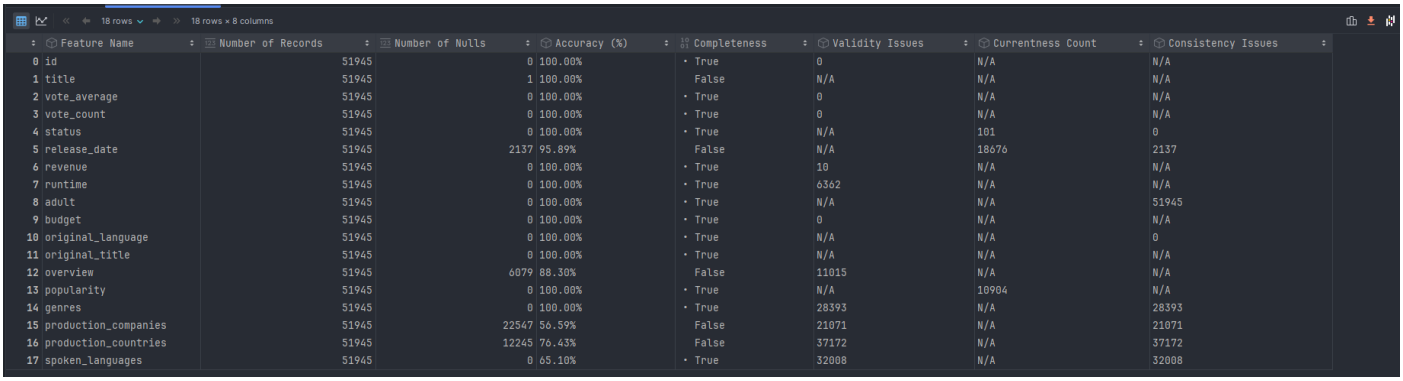
تحلیل خروجی نمودارهای جعبه‌ای:

1. **vote_average**:
   - این ویژگی داده پرت ندارد و مقادیر در محدوده‌ی مشخصی توزیع شده‌اند.
2. **vote_count**:
   - تعداد زیادی داده پرت دارد که نشان‌دهنده‌ی فیلم‌هایی با تعداد رأی بسیار بالا یا پایین است.
3. **runtime**:
   - تعداد قابل توجهی داده پرت نشان می‌دهد که برخی فیلم‌ها دارای زمان پخش غیرعادی هستند.
4. **revenue** و **budget**:
   - مقادیر پرت قابل توجهی در هر دو ویژگی مشاهده می‌شود که ممکن است مربوط به فیلم‌هایی با درآمد یا بودجه بسیار بالا باشد.
5. **popularity**:
   - نمودار جعبه‌ای این ویژگی نیز داده پرت‌های زیادی را نمایش می‌دهد که نشان‌دهنده‌ی محبوبیت غیرعادی برخی از فیلم‌ها است.

نتیجه‌گیری:

نمودارهای جعبه‌ای به شناسایی داده‌های پرت کمک می‌کنند که در تحلیل داده‌ها بسیار مهم هستند. این داده‌های پرت می‌توانند در مراحل بعدی پردازش داده‌ها حذف یا تعدیل شوند تا نتایج بهتری به دست آید.

بخش دوم ارزیابی کیفیت داده ها



	Feature Name	Number of Records	Number of Nulls	Accuracy (%)	Completeness	Validity Issues	Currentness Count	Consistency Issues
0	id	51945	0	100.00%	True	0	N/A	N/A
1	title	51945	1	100.00%	False	N/A	N/A	N/A
2	vote_average	51945	0	100.00%	True	0	N/A	N/A
3	vote_count	51945	0	100.00%	True	0	N/A	N/A
4	status	51945	0	100.00%	True	N/A	101	0
5	release_date	51945	2137	95.89%	False	N/A	18676	2137
6	revenue	51945	0	100.00%	True	10	N/A	N/A
7	runtime	51945	0	100.00%	True	6362	N/A	N/A
8	adult	51945	0	100.00%	True	N/A	N/A	51945
9	budget	51945	0	100.00%	True	0	N/A	N/A
10	original_language	51945	0	100.00%	True	N/A	N/A	0
11	original_title	51945	0	100.00%	True	N/A	N/A	N/A
12	overview	51945	6079	88.30%	False	11015	N/A	N/A
13	popularity	51945	0	100.00%	True	N/A	10904	N/A
14	genres	51945	0	100.00%	True	28393	N/A	28393
15	production_companies	51945	22547	56.59%	False	21071	N/A	21071
16	production_countries	51945	12245	76.43%	False	37172	N/A	37172
17	spoken_languages	51945	0	65.10%	True	32008	N/A	32008

جدول ارزیابی کیفیت داده‌ها که توسط کد تولید شده است، نشان‌دهنده‌ی وضعیت کامل هر ویژگی از نظر چهار معیار اصلی است: کامل بودن، دقت، معتبر بودن و سازگاری. این جدول به تحلیل وضعیت داده‌های مختلف کمک می‌کند تا بتوان بهبودهای لازم برای پردازش داده‌ها را انجام داد. در ادامه توضیحات دقیق‌تری برای هر معیار آورده شده است:

تفسیر معیارهای اصلی:

1. **کامل بودن:(Completeness)**
   - این معیار نشان می‌دهد که آیا تمام مقادیر یک ویژگی پر شده‌اند یا خیر. اگر ستون شامل مقدار تهی باشد، به عنوان ناقص در نظر گرفته می‌شود.
   - به عنوان مثال، release_dateبا تعداد 2137 مقدار تهی و overviewبا تعداد 6079 مقدار تهی، نشان می‌دهند که این ستون‌ها نیاز به تکمیل داده دارند.
2. **میزان دقت:(Accuracy)**
   - درصد مقادیر غیر تهی برای هر ویژگی محاسبه شده و به عنوان درصد دقت گزارش می‌شود. ستون‌هایی با درصد بالاتر از 95٪ معمولاً مناسب هستند.
   - spoken_languagesبا 65.10٪ یکی از ستون‌هایی است که میزان دقت کمتری دارد.
3. **مسائل معتبر بودن:(Validity Issues)**
   - این معیار مواردی را شناسایی می‌کند که با مقدارهای منطقی مغایرت دارند، مانند اعداد غیر معتبر یا مقادیر منفی.
   - runtimeبا 6362 مورد نادرست و genresبا 28393 داده نامعتبر، نمونه‌هایی از مشکلات معتبر بودن هستند.
4. **مسائل سازگاری:(Consistency Issues)**
   - این معیار به بررسی مقادیر ناسازگار می‌پردازد که با استانداردهای مورد انتظار همخوانی ندارند.
   - ستون‌هایی مانند production_companiesو spoken_languagesبا بیش از 20000 مقدار ناسازگار، نیاز به بررسی بیشتر دارند.

نتایج و بهبودها:

- ستون‌هایی که درصد کامل بودن پایینی دارند، مانند production_companiesو production_countries، نیاز به جمع‌آوری داده‌های تکمیلی یا استراتژی‌هایی برای پر کردن مقادیر تهی دارند.

- داده‌های پرت (outliers) که در ستون‌هایی مانند revenue و runtime وجود دارند، باید به دقت بررسی شده و در صورت لزوم حذف یا تعدیل شوند.
- بررسی مسائل معتبر بودن برای ستون‌هایی مانند genres و spoken_languages نشان می‌دهد که باید بازبینی شود تا فرمت داده‌ها با استانداردهای تعریف شده همخوانی داشته باشند.

این جدول و تحلیل‌ها می‌توانند به عنوان یک گزارش اولیه برای تصمیم‌گیری در مورد پاک‌سازی و پیش‌پردازش داده‌ها در پروژه‌های داده‌کاوی یا تحلیل داده مورد استفاده قرار گیرند.

Single-Schema Errors:

این نوع خطاها به مشکلاتی اشاره دارند که به علت عدم تطابق ساختار داده‌ها با قالب تعریف شده یا استاندارد مورد انتظار رخ می‌دهند. به عبارتی دیگر، خطاهایی که از نظر ساختار داده‌ها (مانند نوع داده، قالب‌بندی، و تعریفات اولیه ستون‌ها) به وجود می‌آیند.

موارد مرتبط با خطاهای Single-Schema در این جدول شامل:

1. Consistency Issues:

- ستون original_language: وجود خطا در کد زبان اصلی فیلم‌ها. احتمالاً برخی از داده‌ها در این ستون قالب دو حرفی را رعایت نکرده‌اند.

- ستون status: خطاهایی که ناشی از وضعیت نادرست فیلم‌ها می‌باشند (مواردی که به جای استفاده از وضعیت‌های تعریف‌شده، از موارد دیگری استفاده شده است).

- ستون genres, production_companies, production_countries, spoken_languages: این ستون‌ها شامل داده‌هایی هستند که به جای اینکه لیست یا کاما باشند، به شکل نادرستی ذخیره شده‌اند.

2. Validity Issues:

- ستون‌هایی که اعداد خارج از محدوده دارند، مانند vote_average (باید بین ۰ تا ۱۰ باشد)، vote_count (باید عدد صحیح و غیر منفی باشد)، budget, revenue (نباید منفی باشند) و runtime (باید بین ۰ تا ۳۰۰ دقیقه باشد).

- ستون imdb_id: برخی از آی‌دی‌ها فرمت صحیح `ttXXXXXXX` را رعایت نکرده‌اند.

Single-Instance Errors:

این نوع خطاها به مشکلاتی اشاره دارند که به علت داده‌های نادرست یا غیرمنطقی در سطح رکوردها به وجود آمده‌اند. به عبارتی دیگر، خطاهایی که به دلیل کیفیت پایین یا نبود داده‌های صحیح در برخی از رکوردهای منفرد رخ می‌دهند.

موارد مرتبط با خطاهای Single-Instance در این جدول شامل:

1. Completeness:

- ستون‌های release_date, overview, genres, production_companies و production_countries درصدی از داده‌های خود را از دست داده‌اند و دارای مقدار NULL هستند.

2. Currentness Count:

- ستون `release_date` و `status` بررسی شده‌اند که آیا داده‌های مربوط به تاریخ انتشار با وضعیت فیلم همخوانی دارند یا خیر.

- ستون `popularity`: داده‌های این ستون ممکن است قدیمی باشند و نیاز به به‌روزرسانی دارند تا بازتاب‌دهنده وضعیت فعلی محبوبیت فیلم‌ها باشند.

3. Accuracy:

- برای محاسبه صحت داده‌ها، برخی ستون‌ها مثل `genres`, `production_companies`, `production_countries` دقت پایینی دارند (زیرا با مقدار مرجع خود همخوانی ندارند).

با استفاده از این اطلاعات، می‌توان موارد مشکل‌ساز هر ستون را بر اساس نیازهای کیفیت داده تشخیص داد و در دو ستون Single-Schema و Single-Instance به شکل خلاصه ارائه کرد.

برای بهبود کیفیت داده در این دیتاست و رفع مشکلات موجود، راهکارهای زیر را پیشنهاد می‌کنم. این راهکارها شامل روش‌های پیشنهادی برای پر کردن یا جایگزینی مقادیر نامعتبر نیز هستند.

۱. بهبود Consistency

- **release_date**: تاریخ‌هایی که نامعتبر هستند (تاریخ‌هایی که به درستی ثبت نشده‌اند) باید با تاریخ‌های معتبر جایگزین شوند. برای پر کردن مقادیر خالی می‌توان از روش‌های زیر استفاده کرد:
  - **استفاده از تاریخ تخمینی**: اگر اطلاعاتی در مورد سال تولید فیلم داریم، می‌توانیم تاریخ‌های خالی را با اولین روز سال آن تاریخ جایگزین کنیم.
  - **مراجعه به منابع خارجی**: برای تکمیل تاریخ‌های خالی یا نامعتبر می‌توان از پایگاه‌های داده معتبری مثل IMDb استفاده کرد.
- **original_language**: کد زبان اصلی باید در قالب دو حرفی باشد. برای رفع خطاهای این ستون:
  - **کدهای زبان استاندارد**: داده‌های نامعتبر را می‌توان با کدهای استاندارد زبان جایگزین کرد. برای این کار می‌توان به اطلاعات فیلم یا منبع خارجی مراجعه کرد.
  - **تغییر به حالت Null**: در صورتی که زبان فیلم مشخص نیست، مقدار این فیلد را خالی (Null) قرار دهیم.
- **status**: مقادیر وضعیت غیرمعتبر را باید با وضعیت‌های استانداردی مانند Released, In Production, Canceled, و غیره جایگزین کنیم.
  - **پیشنهاد پیش‌فرض**: در صورتی که هیچ وضعیتی مشخص نیست، می‌توان Unknown یا Planned را به عنوان مقدار پیش‌فرض انتخاب کرد.
- **genres, production_companies, production_countries, spoken_languages**: باید بررسی شود که آیا داده‌ها در قالب لیست کاما هستند. اگر نیستند، این موارد را به صورت لیستی معتبر ثبت کنیم.
  - **مراجعه به منبع معتبر**: در صورت دسترسی، اطلاعات مربوط به ژانرها، شرکت‌های تولید و غیره را از منابع خارجی جمع‌آوری و تصحیح کنیم.

۲. بهبود Currentness

- **release_date** و **status**: باید اطمینان حاصل کنیم که وضعیت فیلم‌ها با تاریخ انتشار همخوانی داشته باشد.

- **بروزرسانی وضعیت** :اگر فیلمی با وضعیت In Productionدارای تاریخ انتشار در گذشته است، وضعیت آن را به Releasedتغییر دهیم. همچنین اگر فیلمی دارای تاریخ انتشار آینده است، وضعیت آن باید Plannedیا In Productionباشد.

- **popularity**:مقادیر مربوط به محبوبیت باید به‌روزرسانی شوند.
  - **جمع‌آوری داده‌های به‌روز** :در صورت دسترسی به اطلاعات جدید از منابع خارجی، مقادیر محبوبیت را بر اساس جدیدترین داده‌ها به‌روزرسانی کنیم.

۳ .بهبود Validity

- **id**:برای اطمینان از یکتایی این ستون، باید رکوردهایی که شناسه تکراری دارند را حذف یا ادغام کنیم.
- **vote_average**و **vote_count**: باید اطمینان حاصل کنیم که امتیازات در محدوده معتبر قرار دارند.
  - **اصلاح مقادیر خارج از محدوده** :مقادیر خارج از محدوده را می‌توان به نزدیک‌ترین مقدار معتبر (۰ یا ۱۰) تغییر داد.
- **budget, revenue**:مقادیر منفی باید به صفر یا Null تغییر کنند.
  - **پیش‌فرض صفر** :در صورت عدم دسترسی به اطلاعات معتبر برای این ستون‌ها، می‌توان بودجه و درآمد را صفر یا خالی (Null)قرار داد.
- **runtime**:فیلم‌ها معمولاً زمان نمایش بین ۳۰ تا ۳۰۰ دقیقه دارند.
  - **اصلاح زمان‌های نمایش نامعتبر** :برای فیلم‌هایی که زمان نمایش آن‌ها خارج از این محدوده است، می‌توان مقدار متوسط یا رایج را به عنوان پیش‌فرض جایگزین کرد.
- **imdb_id**:شناسه‌های نامعتبر را باید تصحیح کرد تا با فرمت استاندارد ttXXXXXXXهمخوانی داشته باشد.
  - **مراجعه به IMDb** :در صورت امکان، شناسه‌های نامعتبر را با جستجو در پایگاه‌های داده معتبر اصلاح کنیم.

۴ .بهبود Completeness

- **release_date**:تاریخ‌های خالی را می‌توان با تاریخ تخمینی یا Null جایگزین کرد.
- **overview, genres, production_companies, production_countries**:برای پر کردن داده‌های خالی در این ستون‌ها:
  - **پر کردن با پیش‌فرض** :اگر اطلاعات دقیق وجود ندارد، این مقادیر را به عنوان Unknownیا Not Availableتنظیم کنیم.
  - **استفاده از داده‌های متنی مشابه** :در برخی موارد، از دیگر ستون‌های مشابه می‌توان به عنوان مرجع برای پر کردن مقادیر استفاده کرد.
- **spoken_languages**:در صورتی که اطلاعات دقیقی از زبان فیلم نداریم، می‌توانیم Unknownیا خالی (Null) را به عنوان مقدار پیش‌فرض انتخاب کنیم.

۵ .بهبود Accuracy

- **genres, production_companies, production_countries, spoken_languages**:اطمینان حاصل کنید که مقادیر این ستون‌ها با منابع معتبر همخوانی دارند.

- **جایگزینی با مقادیر استاندارد** :با توجه به منابع خارجی، اطلاعات نادرست را با مقادیر استاندارد جایگزین کنید.

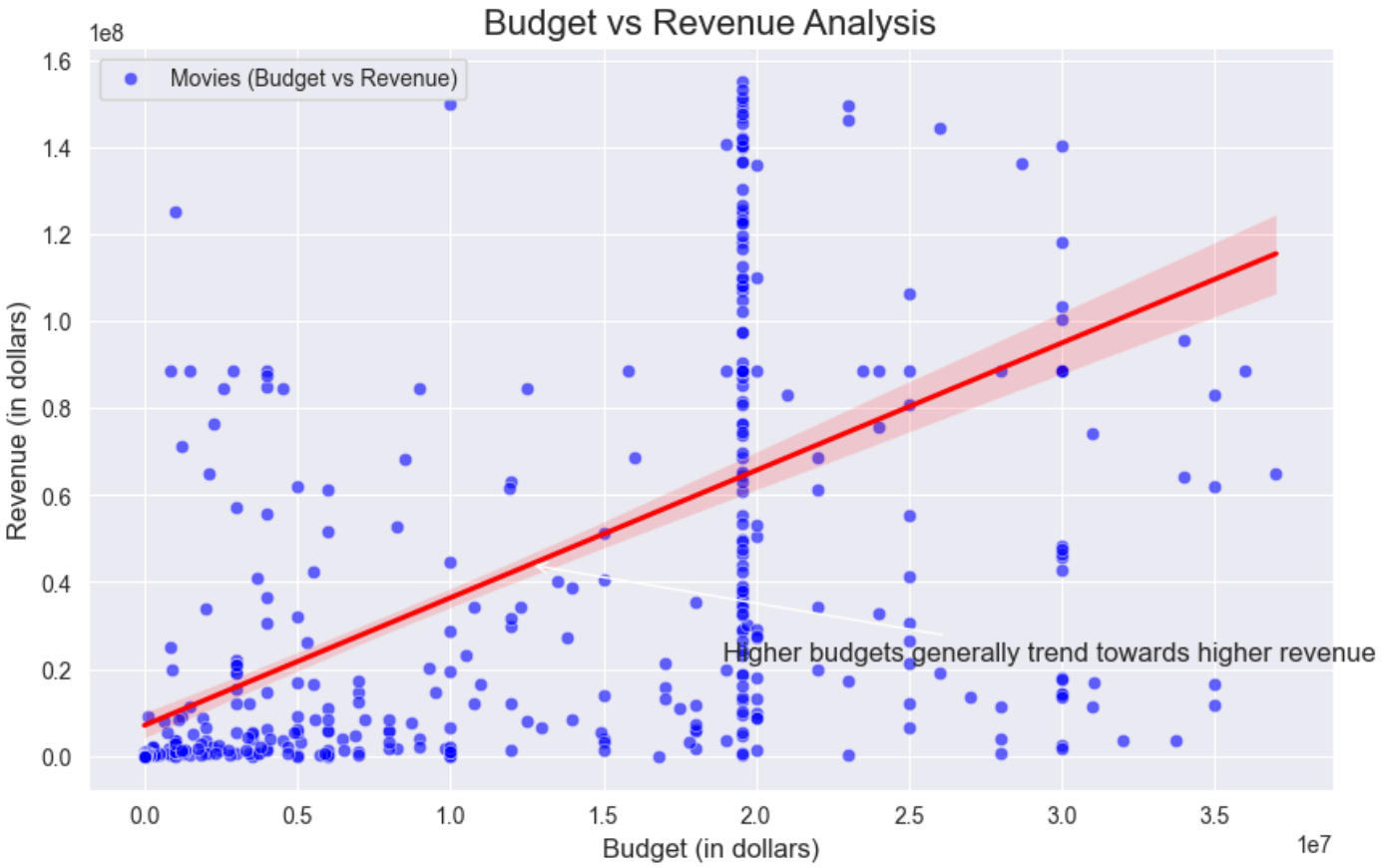
- **title** و **original_title**: برای دقت بیشتر، اطمینان حاصل کنید که این عناوین با منابع خارجی همخوانی دارند. در صورت تفاوت، می‌توانید از عنوان رسمی فیلم استفاده کنید.

خلاصه

با اعمال راهکارهای بالا، کیفیت داده‌ها از نظر کامل بودن، دقت، اعتبار، و بروزرسانی بهبود خواهد یافت. این موارد نه تنها موجب افزایش قابلیت استفاده از دیتاست می‌شوند، بلکه به کاهش خطاها در تحلیل‌ها و نتایج کمک خواهند کرد. برای پر کردن مقادیر خالی، می‌توان از مقادیر پیش‌فرض، منابع خارجی، یا داده‌های مشابه در مجموعه داده استفاده کرد.

بخش چهارم مصور سازی داده ها

Budget and Revenue Scatter plot and association inspection



این نمودار **رابطه بین بودجه و درآمد فیلم‌ها** را نشان می‌دهد. در این نمودار، محور افقی نمایانگر **بودجه** و محور عمودی نمایانگر **درآمد** است. نقطه‌های آبی نمایانگر فیلم‌ها هستند، و خط قرمز که از میان داده‌ها عبور می‌کند، **روند کلی ارتباط بین بودجه و درآمد** را نمایش می‌دهد.

تحلیل نمودار به این صورت است:

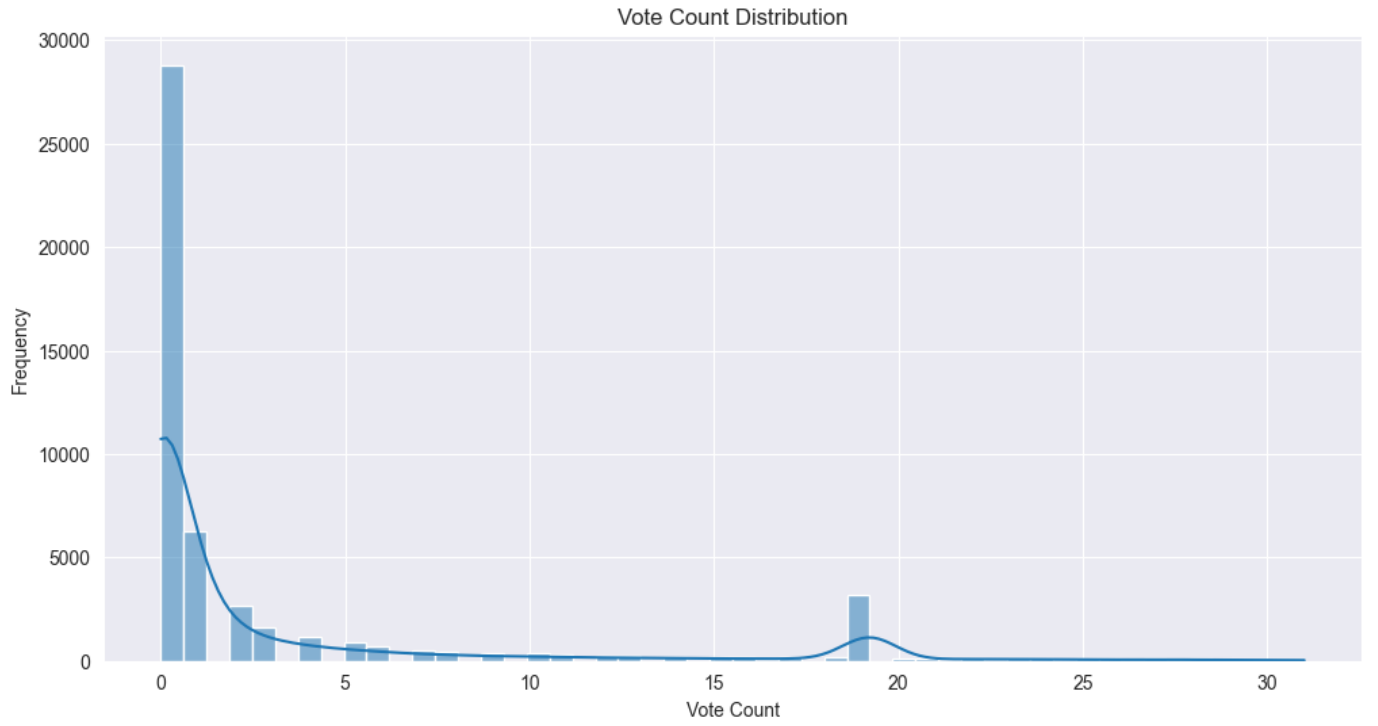
1. **روند مثبت** :خط قرمز به سمت بالا متمایل است که نشان‌دهنده‌ی **رابطه مثبت** بین بودجه و درآمد است؛ یعنی فیلم‌هایی که بودجه بیشتری دارند، به طور کلی درآمد بیشتری کسب کرده‌اند.

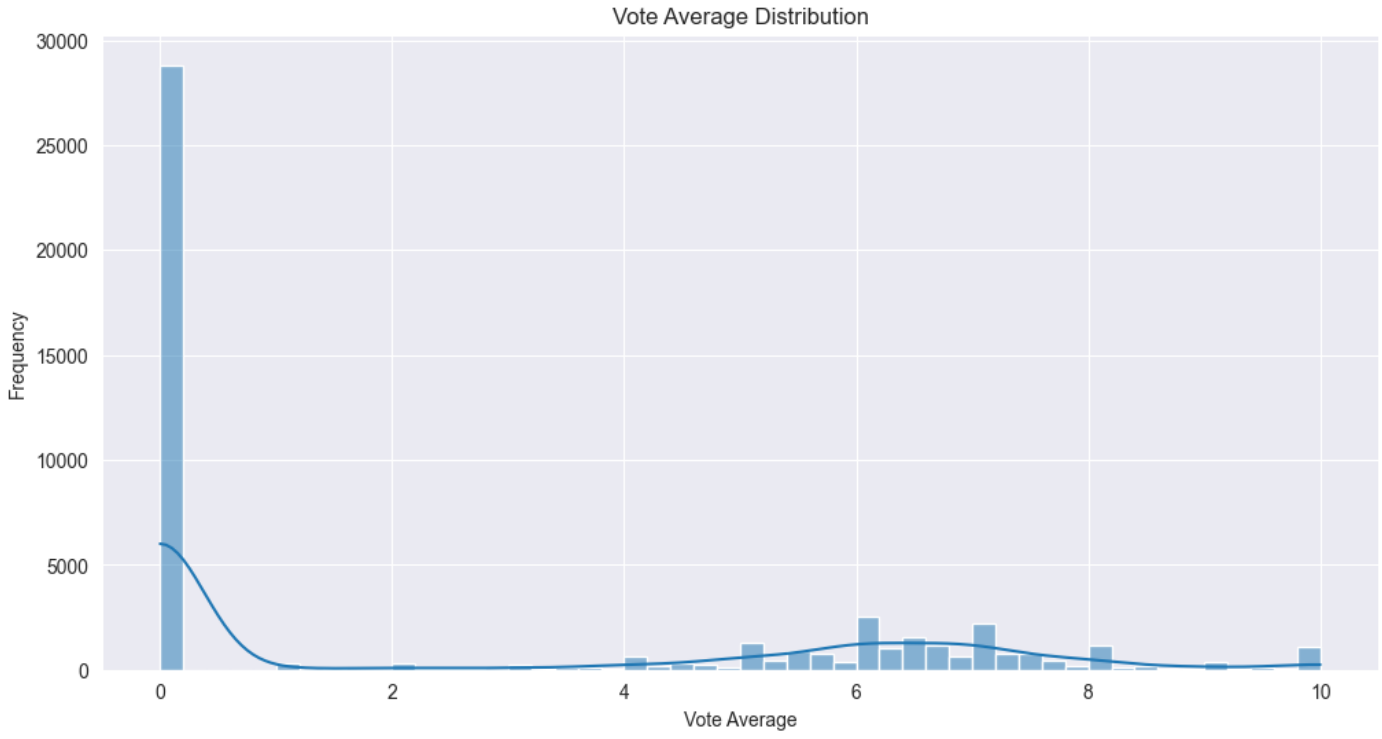
2. **پراکندگی بالا** :در عین حال، پراکندگی بالایی در داده‌ها مشاهده می‌شود. این به این معنی است که حتی فیلم‌هایی با بودجه کمتر هم در برخی موارد توانسته‌اند درآمد قابل توجهی کسب کنند. بنابراین، **بودجه بیشتر لزوماً تضمین‌کننده‌ی درآمد بیشتر نیست** اما احتمال موفقیت بیشتر را افزایش می‌دهد.

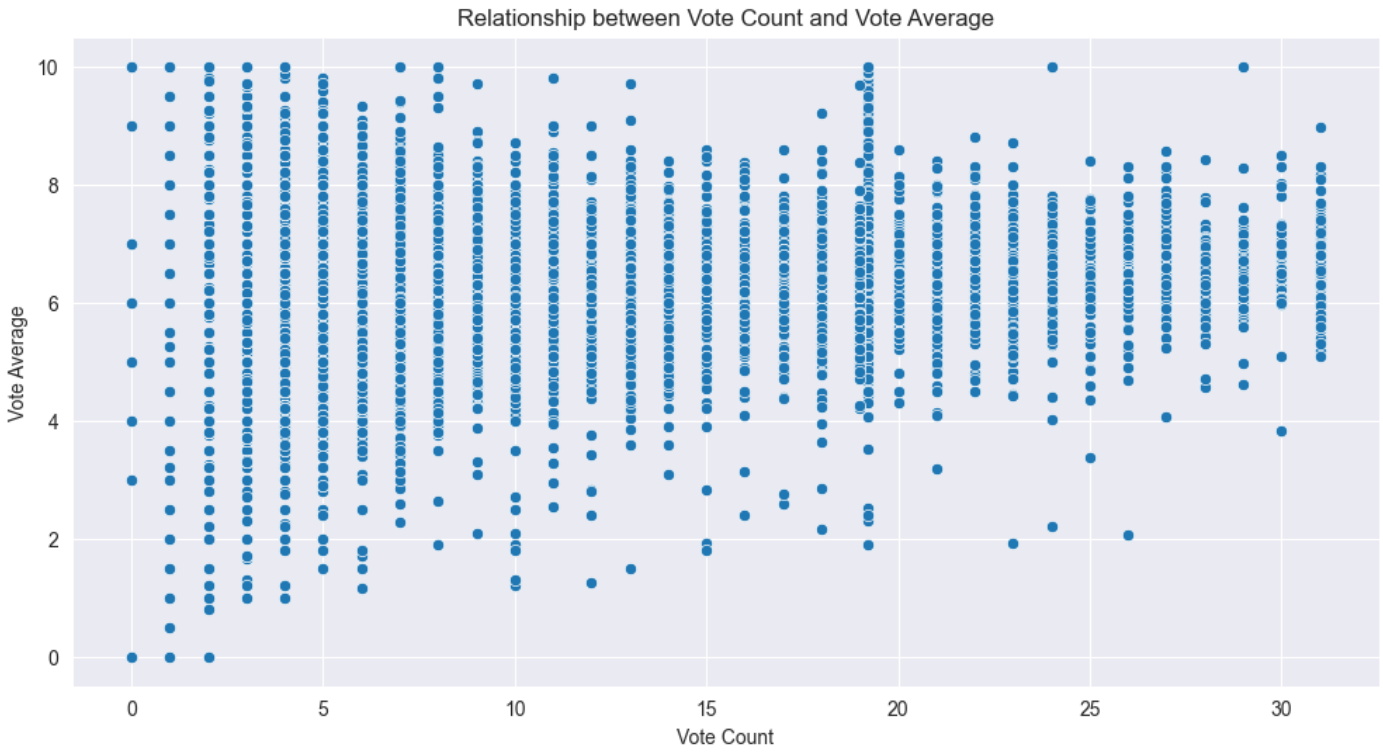
نتیجه‌گیری :به طور کلی، نمودار نشان می‌دهد که **بودجه بیشتر معمولاً به درآمد بیشتر منجر می‌شود**، اما عوامل دیگری نیز تأثیرگذار هستند و تنها بودجه نمی‌تواند به‌طور کامل موفقیت مالی یک فیلم را تضمین کند.

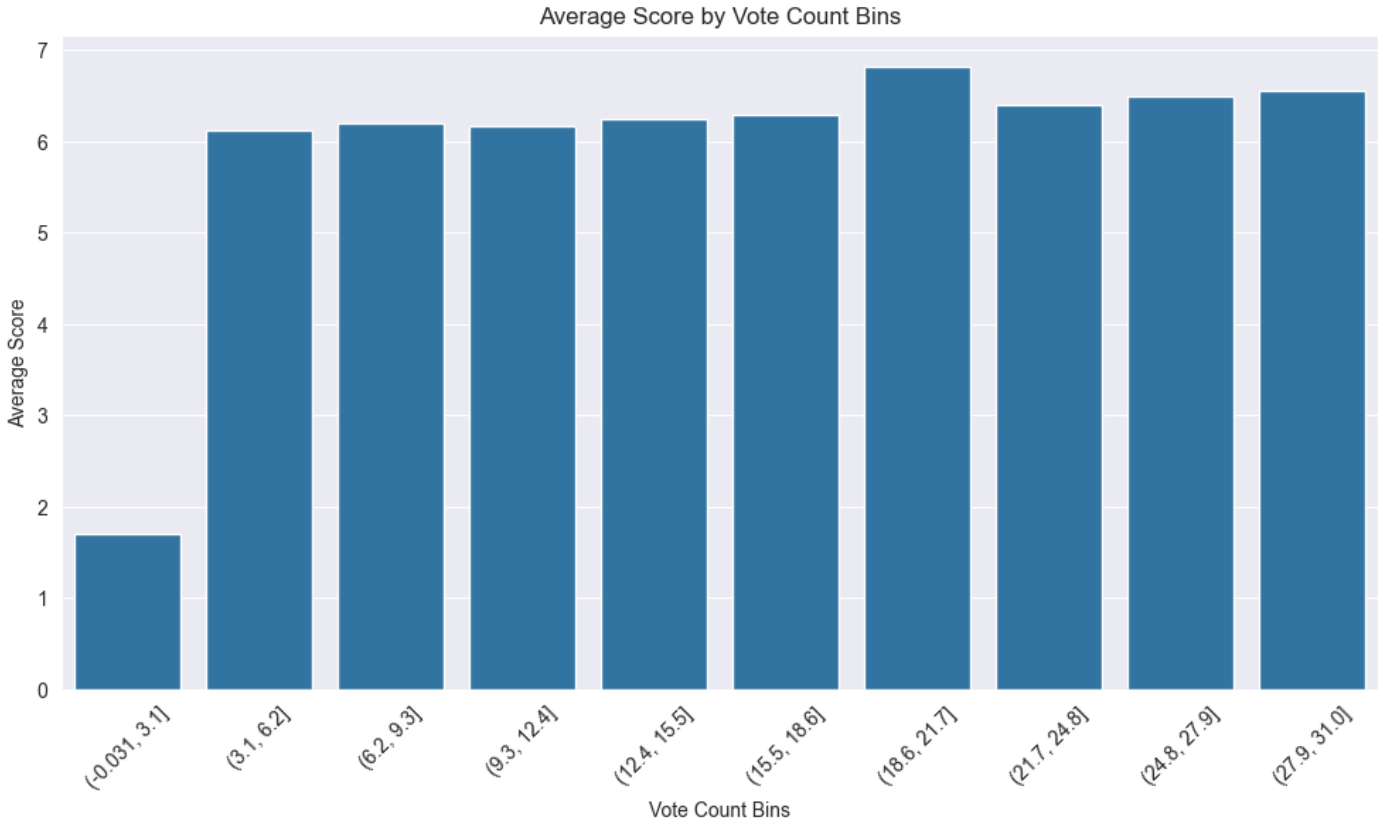
---

vote_average and vote_count histogram inspection









تحلیل جامع بر اساس چهار نمودار:

1. **تعداد رأی و میانگین امتیاز:**

- نمودار اول، "میانگین امتیاز بر اساس بازه‌های تعداد رأی"، یک روند واضح را نشان می‌دهد که در آن اولین بازه (0.03 تا 3.1 رأی) به‌طور قابل‌توجهی میانگین امتیاز پایین‌تری دارد. این امر نشان می‌دهد که فیلم‌هایی با تعداد رأی بسیار کم تمایل به دریافت امتیازات پایین‌تر دارند. با حرکت به سمت بازه‌های با تعداد رأی بیشتر، میانگین امتیازات در حدود 6 تا 7 ثابت می‌شود، که نشان‌دهنده حفظ امتیازات ثابت برای فیلم‌هایی با رأی بیشتر است.

2. **روندهای انتشار فیلم‌های انیمیشنی در طول سال‌ها:**

- نمودار دوم که تعداد فیلم‌های انیمیشنی منتشر شده در هر سال را نشان می‌دهد، افزایش شدیدی از اوایل دهه 2000 را نشان می‌دهد که در حدود سال 2018 به اوج خود می‌رسد. این روند نشان‌دهنده علاقه و سرمایه‌گذاری فزاینده در فیلم‌های انیمیشنی است که با پیشرفت‌های تکنولوژیکی و افزایش تقاضا برای محتوای انیمیشنی هم‌راستا می‌باشد. با این حال، پس از سال 2018 کاهش محسوسی دیده می‌شود که ممکن است به دلیل اشباع بازار یا عوامل خارجی مانند همه‌گیری کووید-19 بر تولید تأثیر گذاشته باشد.

3. **محبوبیت ژانرها:**

- نمودارهای سوم و چهارم ژانرها را بر اساس تعداد فیلم‌ها، میانگین امتیازات رأی‌دهی، و درآمد تحلیل می‌کنند.
- **تعداد فیلم‌ها بر اساس ژانر** :اکشن و کمدی در تولید پیشتاز هستند که نشان‌دهنده محبوبیت آنها در میان فیلم‌سازان و مخاطبان است.
- **میانگین امتیاز رأی‌دهی بر اساس ژانر** :ژانرهای اکشن، تئاتر، و برنده جوایز بالاترین میانگین امتیازات را دارند که نشان‌دهنده تأیید قوی مخاطبان است.
- **میانگین درآمد بر اساس ژانر** :ژانرهای ماجراجویی و خانوادگی بیشترین درآمد را دارند که نشان‌دهنده موفقیت تجاری آنها است.

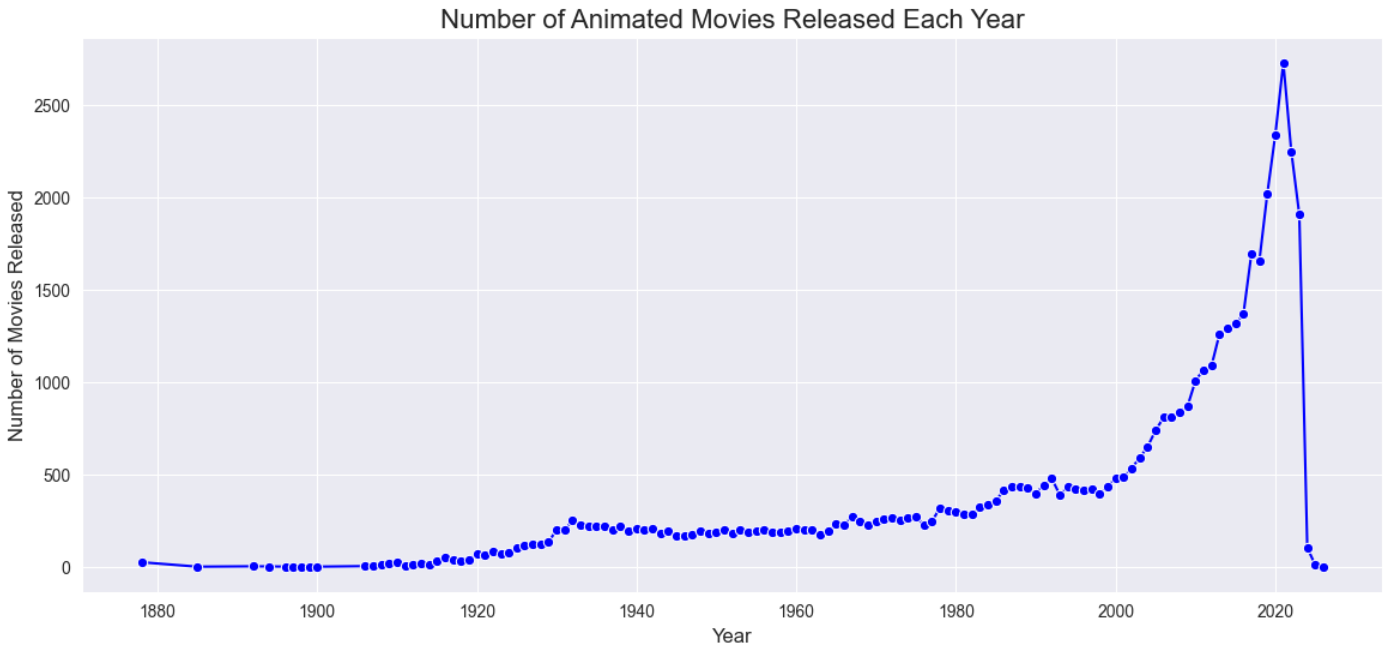
نتیجه‌گیری:

1. **ارتباط تعداد رأی‌های بالا با امتیازات ثابت** :فیلم‌هایی با تعداد رأی‌های بیشتر، تمایل به داشتن میانگین امتیازات پایدار و بالاتری دارند.
2. **افزایش و کاهش در انتشار فیلم‌های انیمیشنی** :فیلم‌های انیمیشنی در اوایل قرن 21 شاهد افزایش شدیدی در انتشار بوده‌اند که در حدود سال 2018 به اوج خود رسیده و پس از آن کاهش یافته‌اند.
3. **ژانرهای محبوب و موفق:**

- **اکشن** :از نظر تعداد فیلم‌ها و امتیاز مخاطبان محبوبیت دارد.
- **ماجراجویی و خانوادگی** :از نظر تجاری بسیار موفق هستند و بالاترین درآمد را دارند.
- **کمدی** :از نظر تعداد فیلم‌ها تولید بالایی دارد و از محبوبیت بالایی برخوردار است.

به‌طور کلی، تحلیل نشان می‌دهد که در حالی که برخی ژانرها مانند اکشن و کمدی به‌طور مداوم محبوب هستند، ژانرهایی مانند ماجراجویی و خانوادگی از نظر موفقیت تجاری برجسته‌اند. روندهای انتشار فیلم‌های انیمیشنی نیز نشان‌دهنده ماهیت پویا صنعت فیلم است که تحت تأثیر عوامل تکنولوژیکی و اجتماعی قرار دارد.

Analyzing the Release Trends of Animated Movies Over the Years



تحلیل نمودار خطی:

مشاهدات:

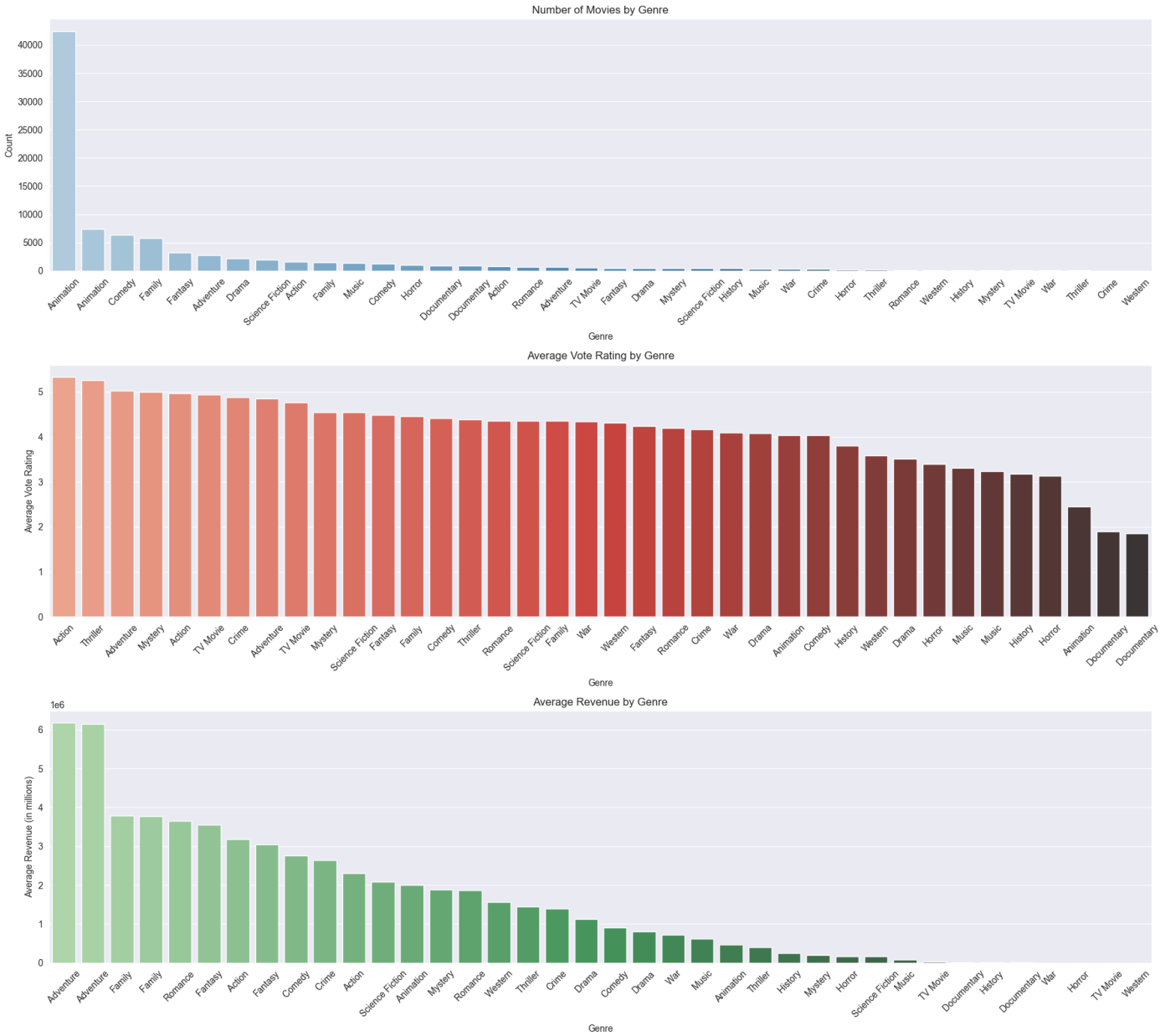
1. **سال‌های ابتدایی ۱۹۲۰-۱۸۸۰:**
   - تعداد فیلم‌های انیمیشنی منتشر شده هر سال بسیار کم بوده و اغلب نزدیک به صفر است.
   - این دوره احتمالاً بازتاب دهنده آغاز کار انیمیشن به عنوان یک رسانه است.
2. **افزایش تدریجی ۱۹۸۰-۱۹۲۰:**
   - یک افزایش تدریجی اما ثابت در تعداد فیلم‌های انیمیشنی منتشر شده مشاهده می‌شود.
   - نوسانات گاه به گاه نشان‌دهنده دوره‌هایی از رشد و رکود است که احتمالاً تحت تأثیر پیشرفت‌های فناوری، شرایط اقتصادی یا تغییرات در تقاضای مخاطبان بوده است.
3. **رشد قابل توجه ۲۰۰۰-۱۹۸۰:**
   - تعداد انتشارها در این دوره به طور قابل توجهی افزایش می‌یابد.
   - روند صعودی ثابت نشان‌دهنده محبوبیت رو به رشد و پیشرفت در فناوری‌های انیمیشن است که تولید فیلم‌های انیمیشنی را آسان‌تر و مقرون به صرفه‌تر کرده است.
4. **افزایش چشمگیر در قرن ۲۱ ۲۰۲۰-۲۰۰۰:**
   - افزایش قابل توجهی در تعداد فیلم‌های انیمیشنی منتشر شده در هر سال دیده می‌شود که در حدود سال ۲۰۱۸ به اوج خود می‌رسد و تعداد فیلم‌ها از ۲۵۰۰ نیز فراتر می‌رود.
   - این افزایش چشمگیر ممکن است به عواملی چون انقلاب دیجیتال، ظهور فناوری CGI (تصاویر تولید شده توسط کامپیوتر)، و افزایش تقاضا برای محتوای انیمیشن از سوی پلتفرم‌های جدید پخش آنلاین مرتبط باشد.
5. **کاهش اخیر (پس از ۲۰۱۸):**

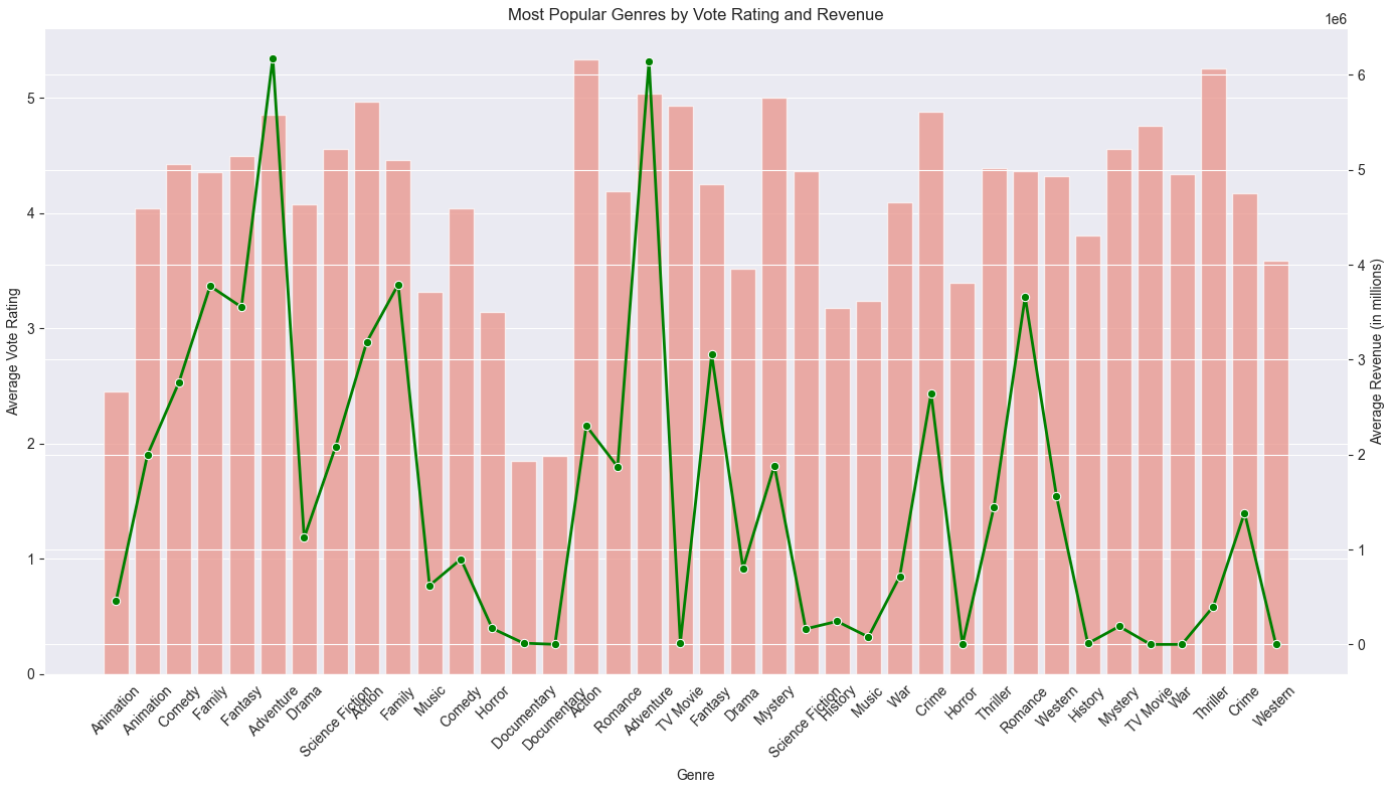
- پس از سال ۲۰۱۸، کاهش شدیدی در تعداد فیلم‌های منتشر شده مشاهده می‌شود.
- این کاهش می‌تواند ناشی از اشباع بازار، تغییر در ترجیحات مصرف‌کنندگان یا عوامل خارجی مانند رویدادهای جهانی) مانند همه‌گیری (COVID-19 باشد که بر تولید و برنامه‌های انتشار تأثیر گذاشته‌اند.

نتیجه‌گیری:

این نمودار به وضوح نشان می‌دهد که تولید فیلم‌های انیمیشنی در طول دهه‌ها به طور تصاعدی رشد کرده، به‌ویژه در قرن ۲۱. اما کاهش شدید اخیر پس از ۲۰۱۸ حاکی از آن است که صنعت انیمیشن ممکن است در حال تجربه دوره‌ای از تعدیل باشد که می‌تواند تحت تأثیر عوامل اقتصادی، فناوری و اجتماعی گوناگون باشد.

Genre Analysis: Identifying Popular and High-Performing Movie Genres





تحلیل از سه نمودار اول:

1. **تعداد فیلم‌ها بر اساس ژانر:**
   - **اکشن** :این ژانر بیشترین تعداد فیلم‌ها را دارد که نشان‌دهنده محبوبیت آن در تولید است.
   - **کمدی** و **خانوادگی** :این ژانرها نیز تعداد بالایی از فیلم‌ها را دارند که نشانگر محبوبیت آنها است.
2. **میانگین امتیاز رأی‌دهی بر اساس ژانر:**
   - **اکشن** :بالاترین میانگین امتیاز رأی را دارد که نشان می‌دهد مخاطبان از فیلم‌های این ژانر لذت می‌برند.
   - **تئاتر** و **جوایز برنده‌شده** :این ژانرها نیز امتیازات بالایی دریافت کرده‌اند.
3. **میانگین درآمد بر اساس ژانر:**
   - **ماجراجویی** :در میانگین درآمد پیشتاز است که نشان از عملکرد قوی آن در گیشه دارد.
   - **خانوادگی** و **فانتزی** :این ژانرها نیز دارای درآمد بالایی هستند که نشان‌دهنده موفقیت مالی آنها است.

تحلیل از نمودار چهارم:

1. **بالاترین میانگین امتیاز رأی‌دهی:**
   - **علمی-تخیلی**، **رمانتیک** و **وسترن** :این ژانرها بالاترین میانگین امتیاز رأی را دارند که نشان از رضایت بالای مخاطبان از این ژانرها دارد.
2. **بالاترین میانگین درآمد:**
   - **ماجراجویی**، **فانتزی** و **انیمیشن** :این ژانرها بیشترین میانگین درآمد را دارند که نشان‌دهنده موفقیت تجاری آنها است.

3. **ژانرهای محبوب بر اساس هر دو معیار**:
    - **ماجراجویی** و **فانتزی** :این ژانرها هم از لحاظ محبوبیت در بین مخاطبان و هم از نظر درآمد، موفق هستند و به عنوان ژانرهای محبوب و موفق شناخته می‌شوند.

دیدگاه‌های ترکیبی:

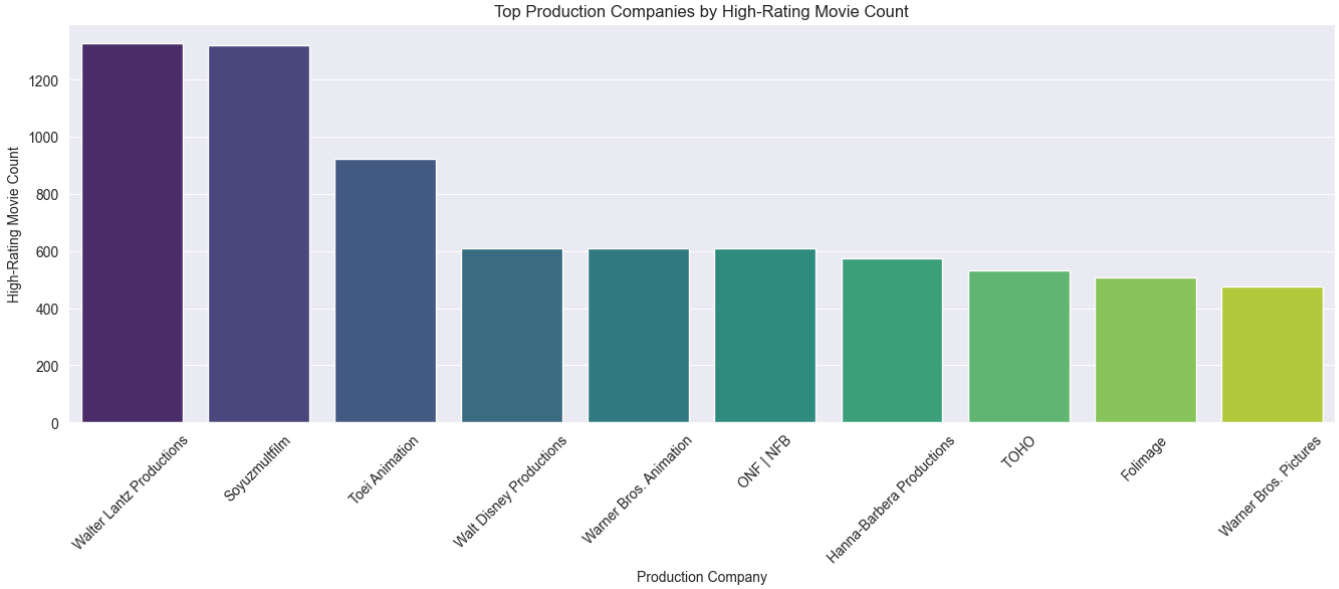
1. **اکشن** :یک ژانر محبوب با تعداد فیلم بالا و بالاترین میانگین امتیاز رأی مخاطبان.
2. **ماجراجویی و فانتزی** :این ژانرها در هر دو زمینه محبوبیت در بین مخاطبان و موفقیت مالی برجسته هستند و به عنوان ژانرهای برتر شناخته می‌شوند.
3. **علمی-تخیلی و رمانتیک** :از سوی مخاطبان امتیاز بالایی دریافت کرده‌اند، هرچند که لزوماً بالاترین درآمد را ندارند.
4. **کمدی و خانوادگی** :محبوبیت بالایی در تولید فیلم‌ها و درآمد دارند.

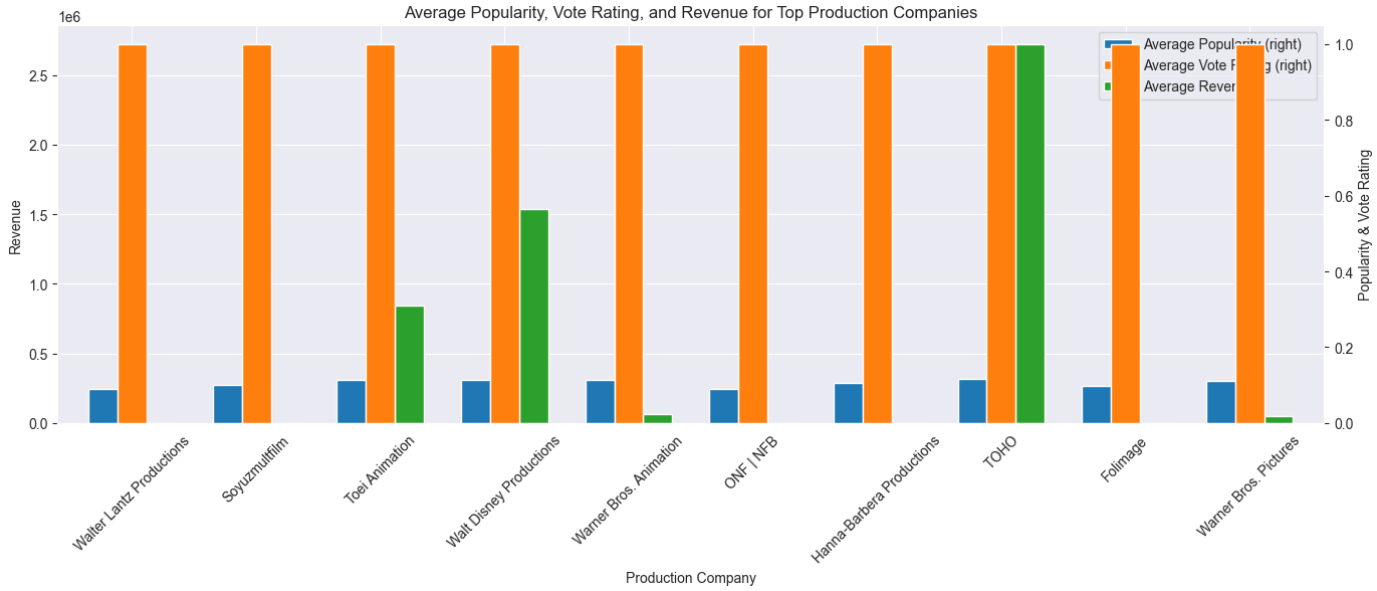
نتیجه‌گیری:

از تحلیل هر دو نمودار، ژانرهای **ماجراجویی** و **فانتزی** به عنوان موفق‌ترین و محبوب‌ترین ژانرها برجسته می‌شوند که دارای امتیازات بالای مخاطبان و درآمد بالایی هستند .**اکشن** نیز از لحاظ تعداد تولید و امتیاز رأی مخاطبان محبوبیت بالایی دارد .**علمی-تخیلی** و **رمانتیک** محبوبیت بالایی از سوی مخاطبان دریافت کرده‌اند، در حالی که **کمدی** و **خانوادگی** به طور مداوم محبوبیت و موفقیت مالی دارند.

این دیدگاه‌ها درک جامعی از محبوبیت و موفقیت ژانرهای مختلف ارائه می‌دهند و نشان می‌دهند کدام ژانرها بر اساس تعداد تولید، امتیاز مخاطبان و درآمد بالاترین جذابیت را دارند.

Analysis of Top Production Companies by High-Rating Movies: Popularity, Average Vote, and Revenue Insights





تحلیل نمودارهای میله‌ای:

تعداد فیلم‌ها بر اساس ژانر:

- ژانر **درام** با بالاترین تعداد فیلم‌ها، پیشرو است.
- ژانرهای **کمدی** و **مستند** در رده‌های بعدی قرار دارند که نشان‌دهنده محبوبیت آنها در تولید فیلم‌ها است.
- ژانرهایی مانند **اکشن**، **ماجراجویی** و **وحشت** نیز تعداد قابل توجهی فیلم دارند که بیانگر جذابیت آنها است.

میانگین امتیاز رأی‌دهی بر اساس ژانر:

- ژانرهای **موسیقی** و **مستند** دارای بالاترین میانگین امتیازات رأی هستند، که نشان می‌دهد فیلم‌های این ژانرها از سوی مخاطبان به خوبی پذیرفته می‌شوند.
- ژانرهای **انیمیشن** و **جنگی** نیز امتیازهای بالایی دارند که بیانگر کیفیت محتوای آنها است.

- ژانرهایی مانند **خانوادگی**، **تاریخی** و **فانتزی** نیز دارای میانگین امتیاز متوسط تا بالا هستند که نشان‌دهنده جذابیت آنها است.

میانگین درآمد بر اساس ژانر:

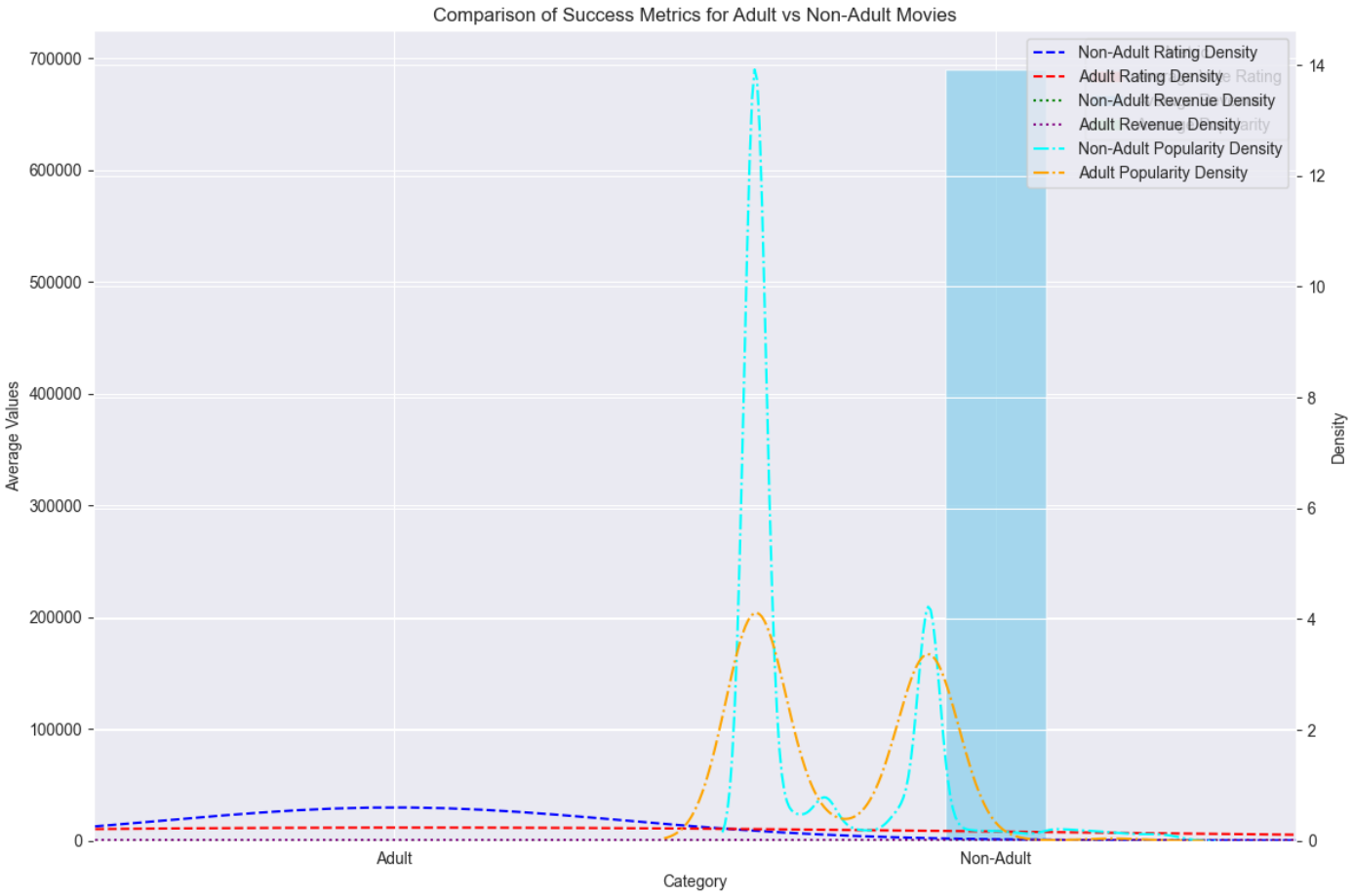
- ژانرهای **انیمیشن** و **خانوادگی** از لحاظ میانگین درآمد پیشتاز هستند که بیانگر موفقیت تجاری آنها است.
- ژانرهای **ماجراجویی** و **فانتزی** نیز دارای میانگین درآمد بالایی هستند که نشان از جذابیت آنها در گیشه دارد.
- سایر ژانرها مانند **علمی-تخیلی** و **اکشن** نیز درآمد قابل توجهی دارند که نشانگر سودآوری آنها است.

نتیجه‌گیری:

از این تحلیل می‌توان نتیجه گرفت که ژانرهایی مانند **درام**، **کمدی** و **مستند** از لحاظ تعداد تولید محبوب هستند. اما ژانرهای **انیمیشن** و **خانوادگی** از نظر موفقیت تجاری برجسته‌اند و بالاترین میانگین درآمد را دارند. همچنین، ژانرهای **موسیقی** و **مستند** از نظر میانگین رأی مخاطبان در صدر هستند.

این نتایج درک جامعی از محبوبیت و موفقیت ژانرهای مختلف فیلم ارائه می‌دهد و به ما کمک می‌کند تا ژانرهایی که بیشترین هماهنگی را با سلیقه مخاطبان دارند و درآمد بالایی تولید می‌کنند، شناسایی کنیم.

Adult-Oriented Movies vs. General Audience Movies: A Success Comparison



این نمودار **مقایسه معیارهای موفقیت فیلم‌های بزرگسال و غیر بزرگسال** را از نظر **امتیاز**، **درآمد**، **و محبوبیت** نشان می‌دهد. در محور افقی، دسته‌بندی فیلم‌ها به دو گروه بزرگسال (Adult) و غیر بزرگسال (Non-Adult) تقسیم شده است، و در محور عمودی، **میانگین معیارهای موفقیت** نمایش داده شده است. هر خط رنگی نمایانگر یک معیار موفقیت برای هر گروه است:

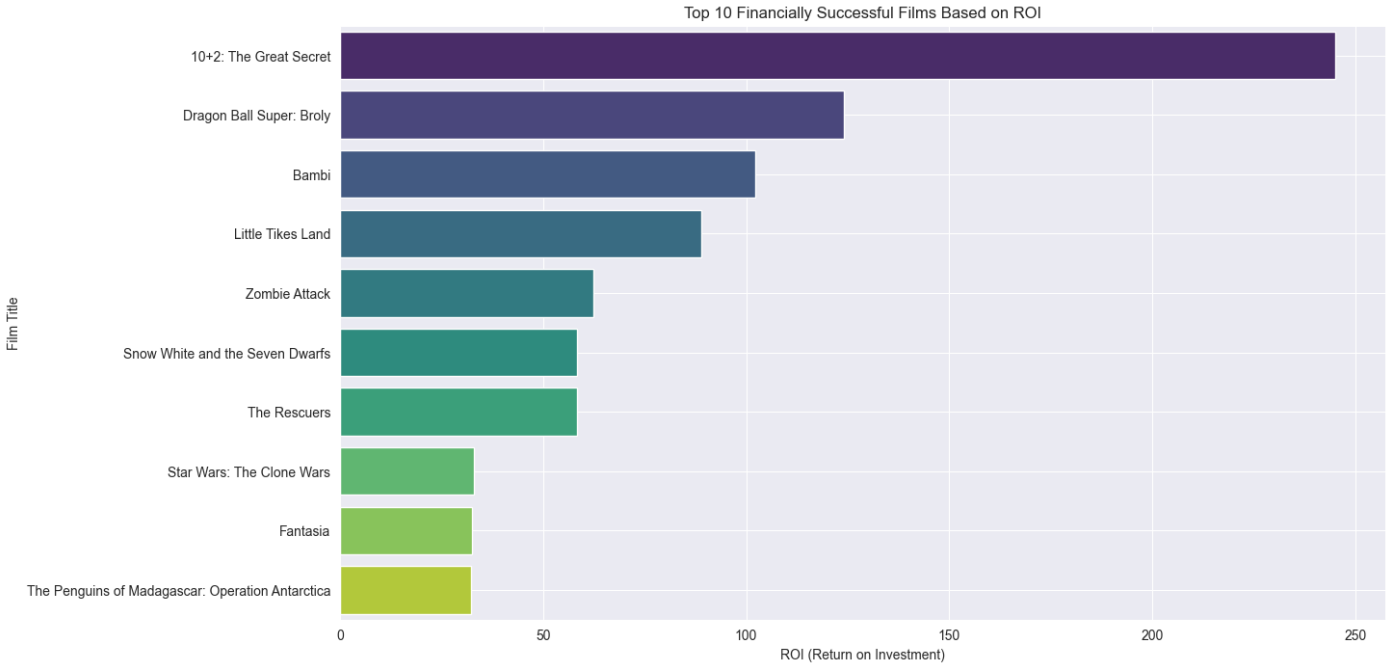
1. **تراکم امتیاز** :خط آبی نقطه‌چین تراکم امتیاز فیلم‌های غیر بزرگسال و خط قرمز نقطه‌چین تراکم امتیاز فیلم‌های بزرگسال را نشان می‌دهد. تراکم امتیاز در هر دو گروه پایین است، ولی تفاوت قابل توجهی بین این دو گروه مشاهده نمی‌شود.

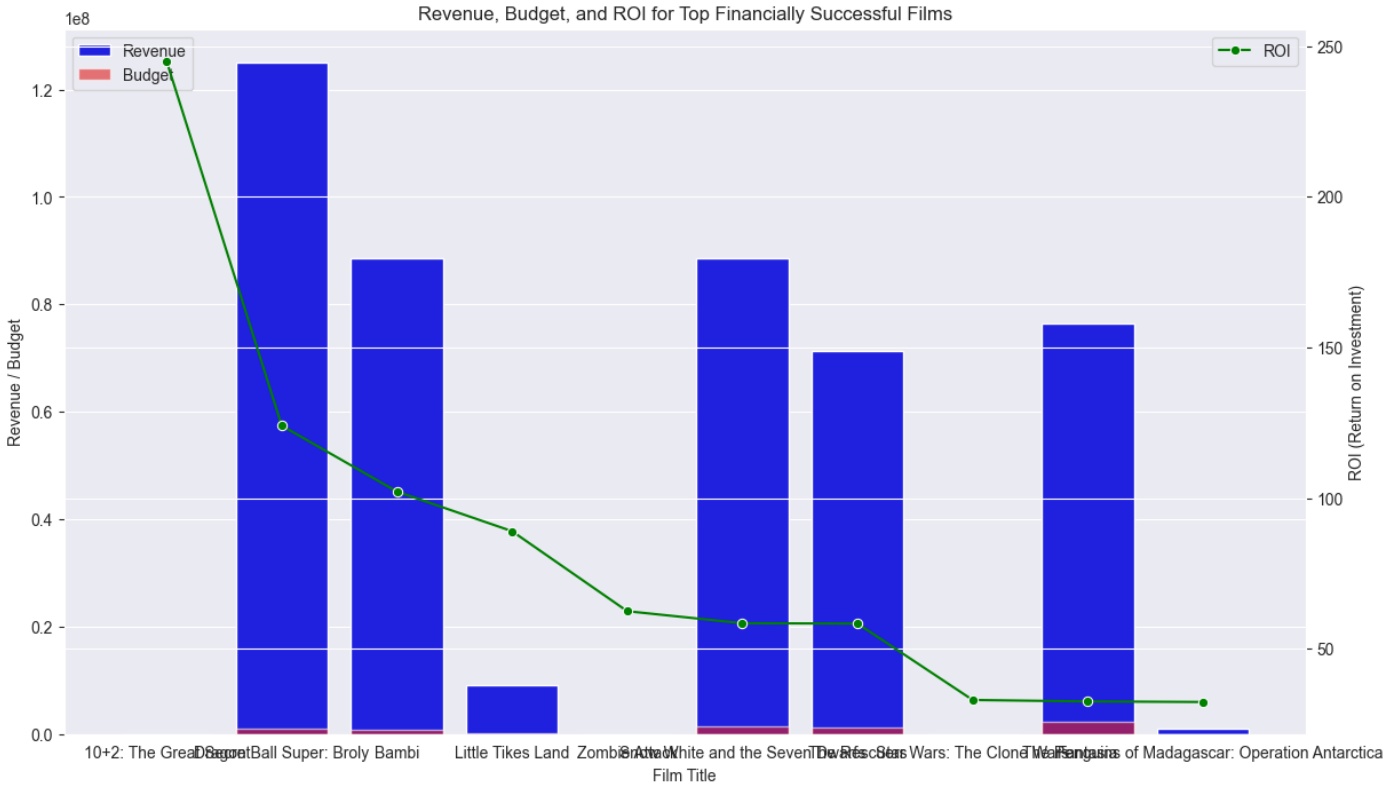
2. **تراکم درآمد** :خط سبز نقطه‌چین برای فیلم‌های غیر بزرگسال و خط بنفش نقطه‌چین برای فیلم‌های بزرگسال تراکم درآمد را نشان می‌دهد. تراکم درآمد برای فیلم‌های بزرگسال نسبت به فیلم‌های غیر بزرگسال بسیار پایین‌تر است، که به این معنی است که **فیلم‌های غیر بزرگسال معمولاً درآمد بیشتری دارند.**

3. **تراکم محبوبیت** :خط نارنجی نقطه‌چین محبوبیت فیلم‌های بزرگسال و خط فیروزه‌ای نقطه‌چین محبوبیت فیلم‌های غیر بزرگسال را نشان می‌دهد. تراکم محبوبیت برای فیلم‌های غیر بزرگسال به مراتب بیشتر است و به ویژه در محدوده‌ای با مقدار بالا متمرکز شده است. این نشان می‌دهد که **فیلم‌های غیر بزرگسال به طور کلی محبوبیت بیشتری دارند.**

نتیجه‌گیری :فیلم‌های غیر بزرگسال به طور کلی **از نظر درآمد و محبوبیت موفق‌تر از فیلم‌های بزرگسال** هستند. با این حال، در زمینه امتیاز تفاوت بارزی وجود ندارد و هر دو گروه تقریباً در سطح مشابهی قرار دارند.

Financial Success Analysis: Revenue-to-Budget Ratio of Movies





در این نمودار ، معیارهای مالی مختلفی برای فیلم‌های موفق از نظر مالی نمایش داده شده است. ستون‌های آبی و قرمز در نمودار به ترتیب نشان‌دهنده **درآمد (Revenue)** و **بودجه (Budget)** هر فیلم هستند، و خط سبز بیانگر **بازده سرمایه‌گذاری (ROI)** است. در جدول نیز اطلاعات **بودجه، درآمد** و **ROI**برای هر فیلم آورده شده است.

تحلیل و نتایج

1. **درآمد نسبت به بودجه:(ROI)**

- فیلم‌ها با ROI بالا نشان‌دهنده موفقیت بیشتری از نظر بازده سرمایه‌گذاری هستند. به عنوان مثال، فیلم "*10+2: The Great Secret*"با ROI حدود 245 و فیلم "*Dragon Ball Super: Broly*" با ROI حدود 124 از نظر مالی بسیار موفق بوده‌اند.
- در مقایسه با این دو فیلم، برخی فیلم‌ها مانند "*The Penguins of Madagascar: Operation Antarctica*" و "*Fantasia*"بازده مالی کمتری داشته‌اند.

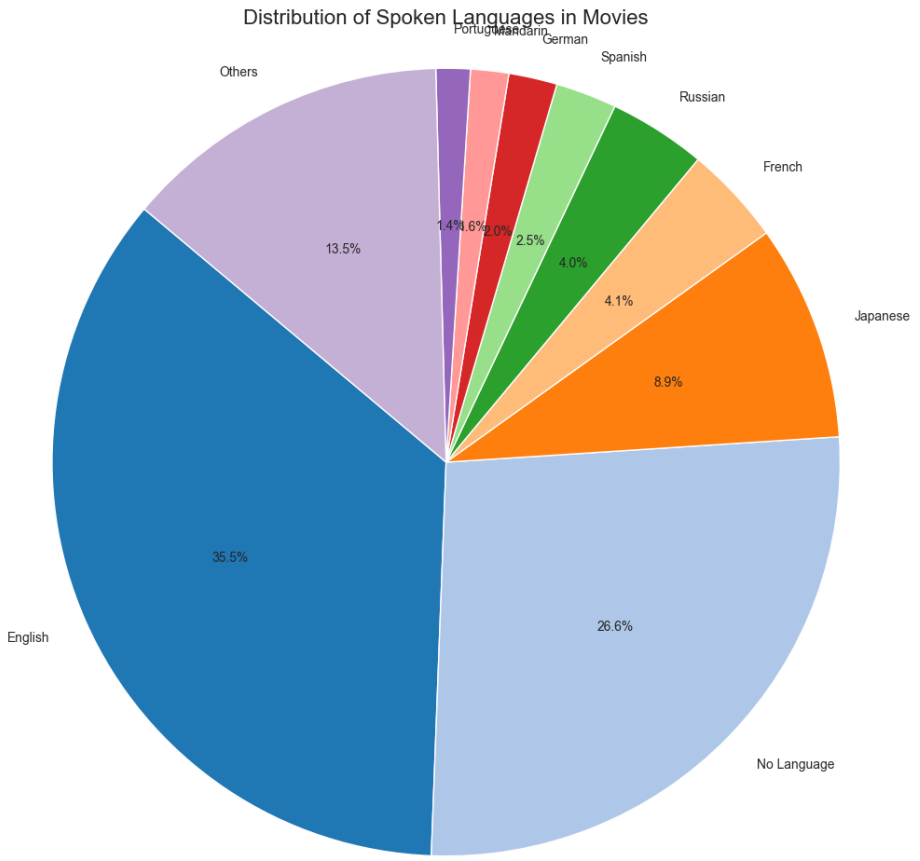
2. **رابطه بین درآمد و بودجه:**

- فیلم‌هایی با بودجه کمتر اما با ROI بالا، معمولاً موفقیت مالی بهتری داشته‌اند، چرا که بازدهی سرمایه بالاتری را به خود اختصاص داده‌اند. برای مثال، فیلم "*Little Tikes Land*" با بودجه کم توانسته ROI قابل توجهی به دست آورد.
- به طور کلی، فیلم‌هایی که بودجه‌های کمتری داشته‌اند در بسیاری از موارد ROI بالاتری داشته‌اند و بنابراین موفقیت مالی بیشتری را از نظر بازده نسبت به سرمایه‌گذاری اولیه نشان داده‌اند.

نتیجه‌گیری

فیلم‌های با بودجه کمتر، اگرچه درآمد کلی کمتری دارند، اما در برخی موارد **ROIبالاتری** نسبت به فیلم‌های با بودجه‌های بالاتر دارند. این موضوع نشان می‌دهد که **موفقیت مالی لزوماً به بودجه زیاد وابسته نیست** و بازدهی سرمایه‌گذاری می‌تواند بسته به شرایط و عملکرد فیلم در بازار، حتی با بودجه‌های کمتر هم بالا باشد.

Language Analysis in Movies: Most Frequently Used Spoken Languages



نتیجه:

این نمودار نشان می‌دهد که زبان انگلیسی بیشترین سهم را در فیلم‌ها دارد و پس از آن فیلم‌های بدون زبان و طیف گسترده‌ای از زبان‌های دیگر قرار دارند. این تنوع، جهانی بودن صنعت فیلم را نشان می‌دهد که در آن زبان‌ها و مناطق مختلف سهم قابل‌توجهی دارند.