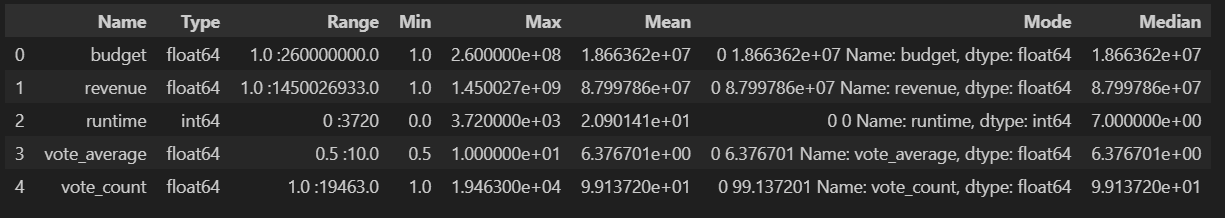
**فاز اول پروژه**

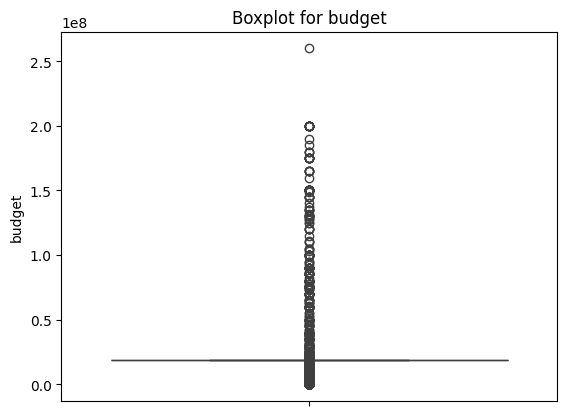
**درس مبانی داده کاوی**

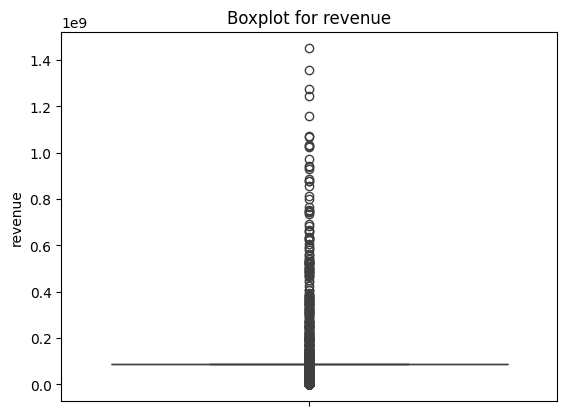
**بخش اول :**

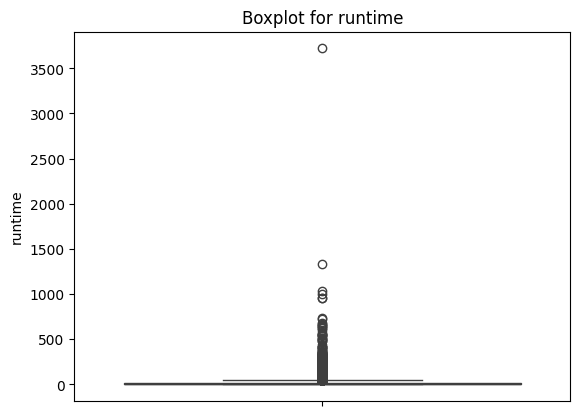
مقادیر اشتباه : ابتدا مقادیر صفر در دیتاست را با میانگین مقادیر دیگر جایگزین میکنیم.

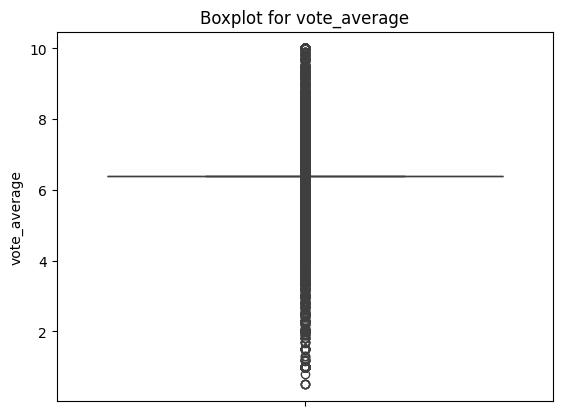


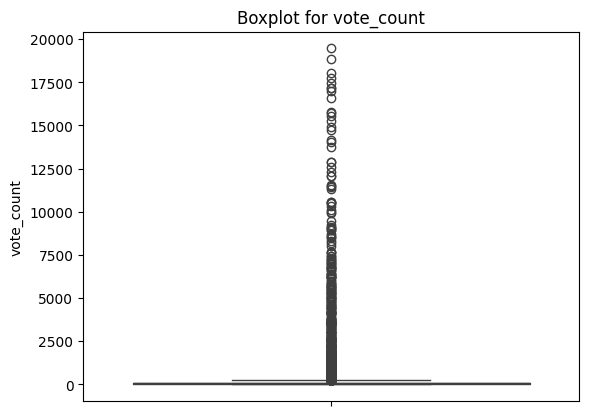
مقادیر دورافتاده : نمودار باکس پلات را برای مقادیر عددی رسم میکنیم.

****









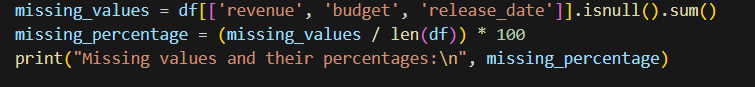
بسیاری از داده ها outlier هستند.

**بخش دوم :**

# تحلیل و بررسی کیفیت داده‌ها

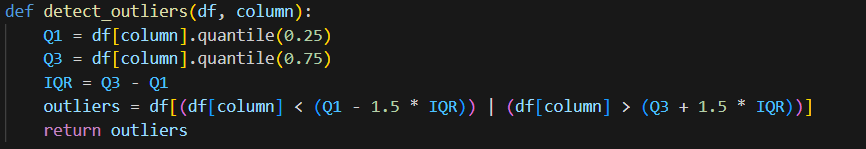
## مرحله ۱: بررسی مقادیر گم‌شده

در این مرحله، مقادیر گم‌شده در ستون‌های بودجه، درآمد و تاریخ انتشار شناسایی می‌شوند. برای این کار، ابتدا تعداد مقادیر گم‌شده در هر ستون محاسبه شده و سپس درصد مقادیر گم‌شده نسبت به کل داده‌ها بدست می‌آید.

کد زیر تعداد و درصد مقادیر گم‌شده را محاسبه و نمایش می‌دهد:  
  


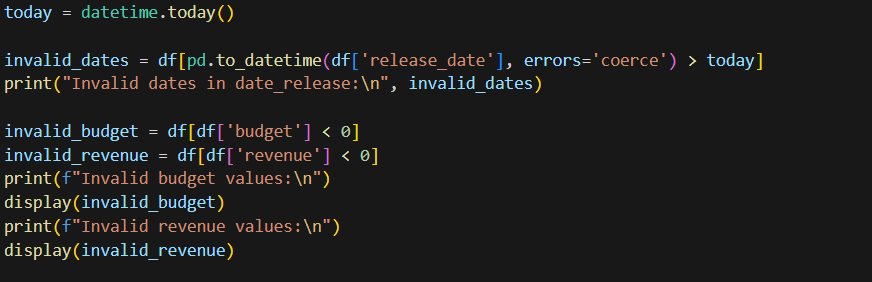
## مرحله ۲: شناسایی داده‌های پرت

داده‌های پرت (outliers) در ستون‌های درآمد و بودجه شناسایی می‌شوند. برای این کار، از شاخص IQR (Interquartile Range) استفاده می‌شود تا مقادیر بسیار دور از حد معمول را تشخیص دهیم.

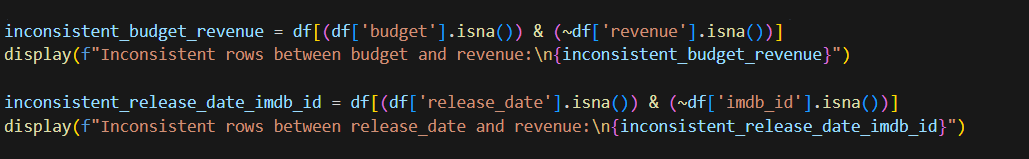
کد زیر برای شناسایی داده‌های پرت در هر ستون نوشته شده است:  
  


## مرحله ۳: بررسی ناسازگاری‌ها و خطاها در تاریخ انتشار

تاریخ‌های انتشار بررسی می‌شوند تا اطمینان حاصل شود که هیچ تاریخ انتشاری در زمان حال یا آینده وجود ندارد. به عنوان مثال، اگر تاریخی در ستون `release\_date` بزرگتر از تاریخ فعلی باشد، آن را نامعتبر در نظر می‌گیریم.

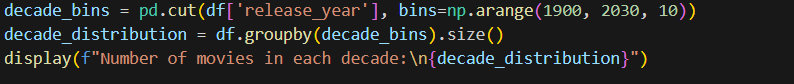
کد زیر برای شناسایی تاریخ‌های نامعتبر و بودجه یا درآمد منفی نوشته شده است:  


## ناسازگاری ها:

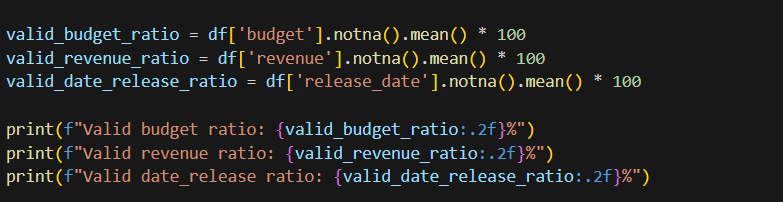
این قسمت از کد به دنبال ردیف‌هایی می‌گردد که در آنها مقدار بودجه (‌budget) موجود نیست ولی مقدار درآمد (‌revenue) موجود است. وجود درآمد بدون مقدار بودجه می‌تواند نشان‌دهنده ناسازگاری یا مشکل در داده‌ها باشد. این ناسازگاری‌ها نمایش داده می‌شوند تا بتوان آن‌ها را بررسی و اصلاح کرد:  


## توزیع فیلم‌ها در دهه‌های مختلف :

در این بخش، فیلم‌ها بر اساس دهه انتشار دسته‌بندی می‌شوند. دهه‌ها به عنوان بازه‌های ۱۰ ساله تعریف می‌شوند و تعداد فیلم‌ها در هر دهه محاسبه می‌شود. این توزیع به نمایش تعداد فیلم‌ها در هر دوره زمانی کمک می‌کند



## محاسبه نسبت مقادیر معتبر در ستون‌های بودجه، درآمد و تاریخ انتشار:

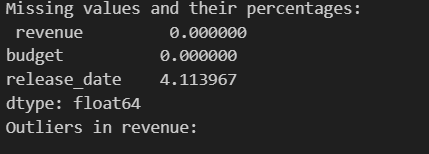
این قسمت از کد درصد مقادیر معتبر (ناموجود نبودن) را برای ستون‌های بودجه، درآمد و تاریخ انتشار محاسبه و نمایش می‌دهد. برای مثال، نسبت معتبر بودجه نشان می‌دهد چند درصد از فیلم‌ها مقدار بودجه دارند. این مقادیر برای ارزیابی کیفیت کلی داده‌ها مهم هستند 

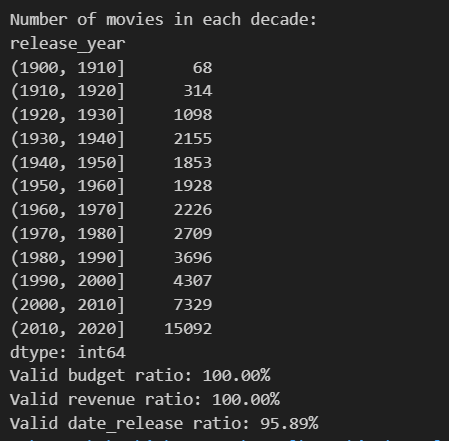
## پیشنهادات

پیشنهادات برای بهبود کیفیت داده‌ها:

* پر کردن مقادیر گم‌شده با میانگین یا میانه‌ی داده‌ها برای ستون‌های budget و revenue
* استفاده از اطلاعات تاریخی معتبر برای پر کردن مقادیر گم‌شده در date\_release
* حذف یا تصحیح مقادیر پرت (outliers) مانند مقادیر منفی در ستون‌های budget وrevenue.

Sبرخی از خروجی های این بخش:

* 

****

**مشکلات مربوط به داده‌ها در جدول می‌توانند شامل مسائل مختلفی باشند که در زمینه‌های زیر خلاصه می‌شوند:**

1. Single Instance (تک نمونه‌ای بودن داده)

- مشکل: داده‌ها تنها شامل یک نمونه از هر نوع (یا کلاس) هستند. این مسئله می‌تواند منجر به مشکل در تحلیل داده‌ها و ایجاد مدل‌های یادگیری ماشین شود، زیرا داده‌های کافی برای شناسایی الگوها و روابط وجود ندارد.

- توضیح: در یادگیری ماشین و تحلیل داده‌ها، دسترسی به داده‌های متنوع و نمونه‌های متعدد از هر نوع ضروری است تا مدل بتواند به درستی تعمیم داده‌ها را یاد بگیرد. تک نمونه‌ای بودن باعث کاهش قدرت تعمیم‌دهی مدل و افزایش خطای آن در پیش‌بینی می‌شود.

2. Single Schema (یک‌دستی یا یکسان بودن ساختار داده)

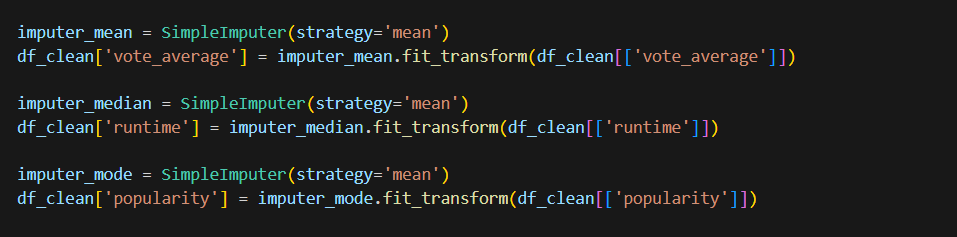
- مشکل: ساختار و فرمت داده‌ها به صورت یکسان است و تفاوتی در میان داده‌های مختلف مشاهده نمی‌شود. این مسئله ممکن است باعث شود که داده‌ها محدودیت در تنوع داشته باشند و برخی از ویژگی‌های لازم را نداشته باشند.

- توضیح: وقتی همه داده‌ها از یک نوع ساختار پیروی می‌کنند، امکان وجود ناهماهنگی و تنوع در داده‌ها کاهش می‌یابد. این مسئله باعث می‌شود که داده‌های مورد نیاز برای یادگیری کامل مدل و پوشش شرایط مختلف موجود نباشند و مدل به داده‌های واقعی قابل تعمیم نباشد. به عنوان مثال، داده‌های مختلف در پایگاه‌های داده‌ای از ساختارهای مختلف استفاده می‌کنند و نیاز است که داده‌ها به شکل استاندارد یا در قالب‌های متنوع‌تری نگهداری شوند.

**بخش سوم:**

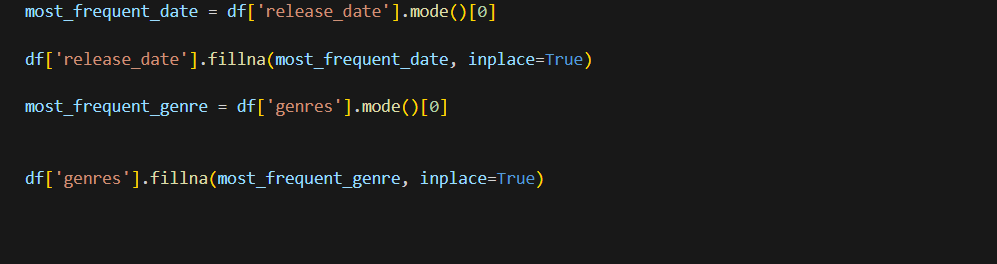
* **پر کردن مقادیر گم‌شده در ستون‌های مختلف با میانگین** :

در اینجا، مقادیر گم‌شده در ستون‌های vote\_average، runtime، و popularity با استفاده از میانگین (mean) آن ستون‌ها جایگزین می‌شوند. استفاده از میانگین برای پر کردن مقادیر گم‌شده عددی می‌تواند به همگن‌سازی داده‌ها کمک کند

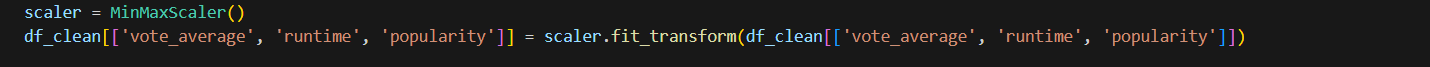
****

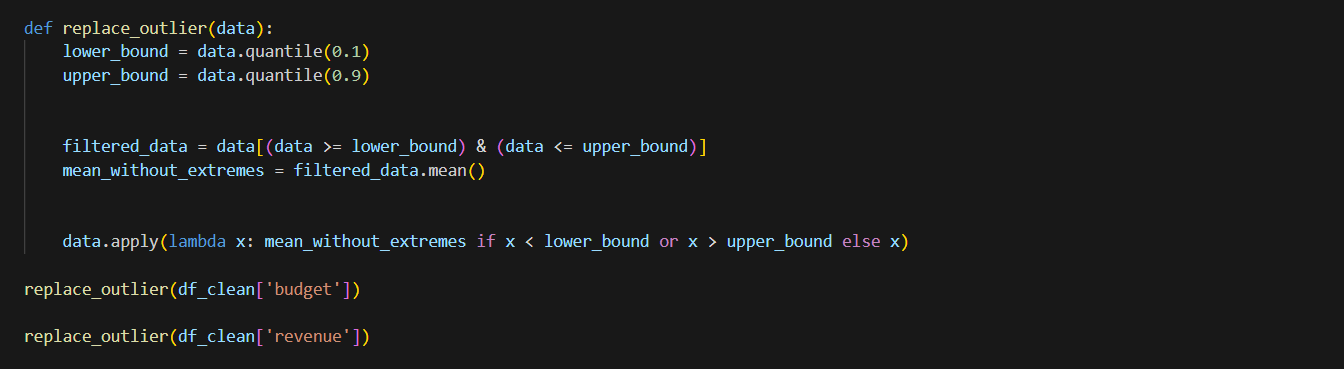
* **جایگزینی مقادیر گم‌شده در ستون‌های تاریخ انتشار و ژانر با مقدار پرتکرار:**

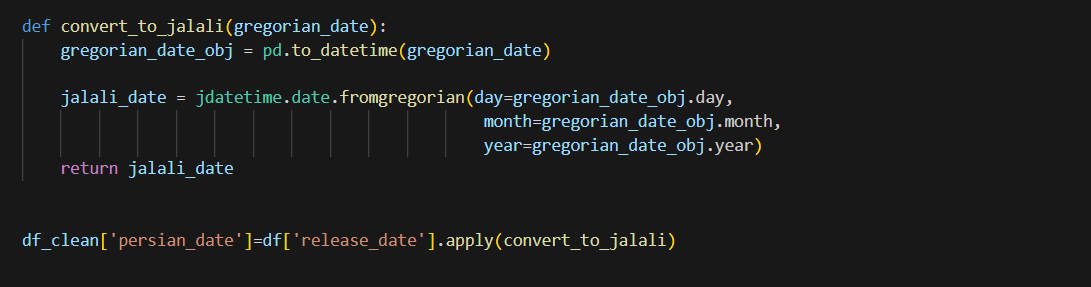
در این بخش، مقادیر گم‌شده در ستون‌های release\_date و genres با مقدار پرتکرار (‌mode) جایگزین می‌شوند. این روش برای ستون‌هایی که داده‌های متنی یا تاریخ هستند مناسب است.

****

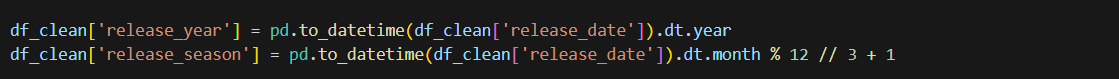
* **نرمال‌سازی ستون‌های عددی:**

****

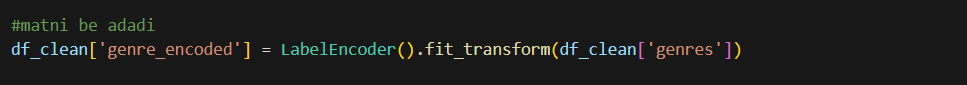
* **جایگزینی داده‌های پرت با میانگین بدون در نظر گرفتن ۱۰٪ بالا و پایین**:
* **ایجاد ستون های جدید:**

****

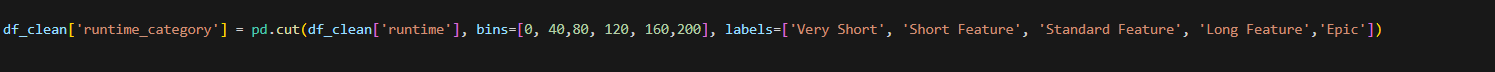
* **استخراج سال و فصل انتشار:**

****

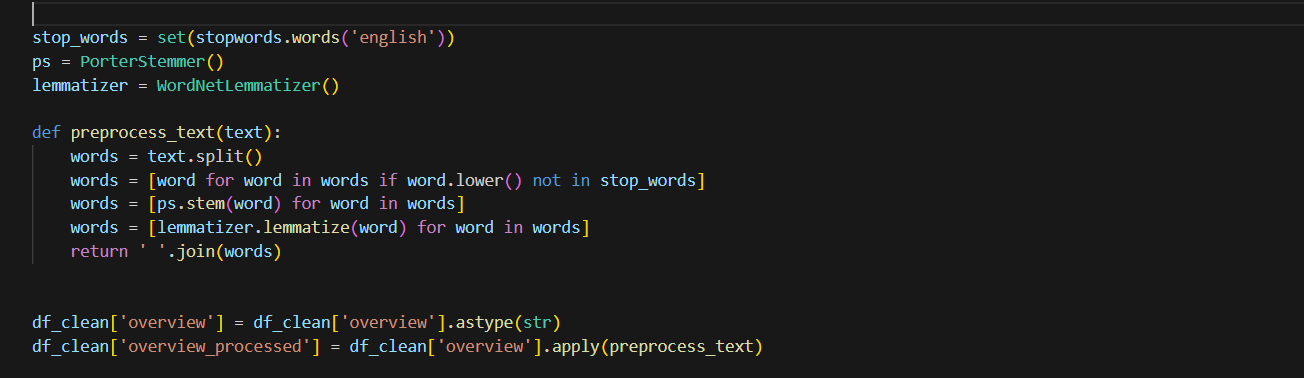
* **رمزگذاری ستون ژانر:**

****این خط ژانرهای مختلف را به مقادیر عددی تبدیل می‌کند تا در تحلیل‌های عددی استفاده شوند.

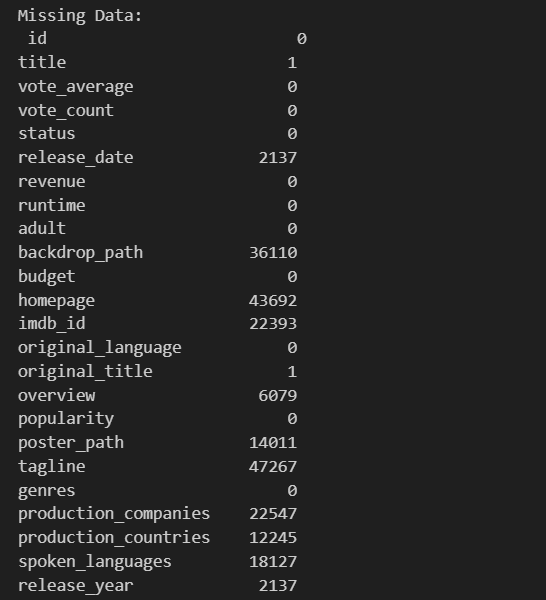
* دسته‌بندی طول فیلم‌ها:



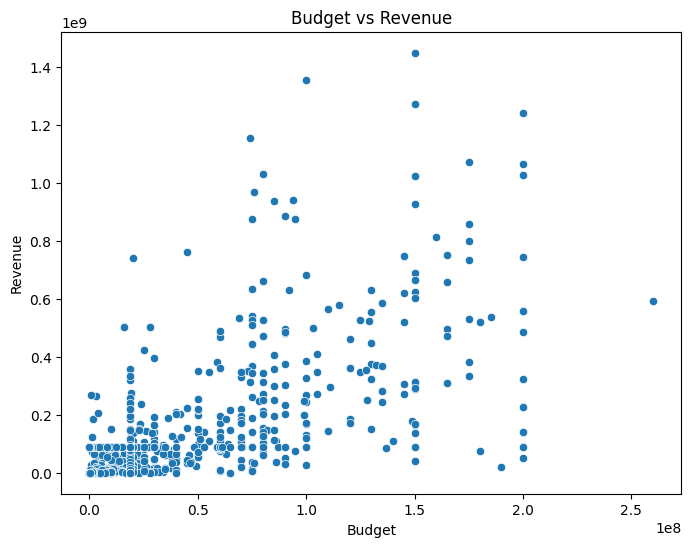
پیش‌پردازش توضیحات (overview) برای تحلیل متن:

****

**خروجی ها :**

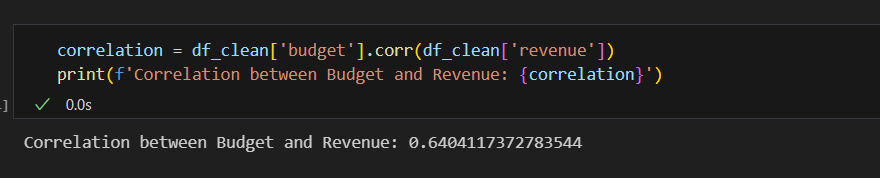
****

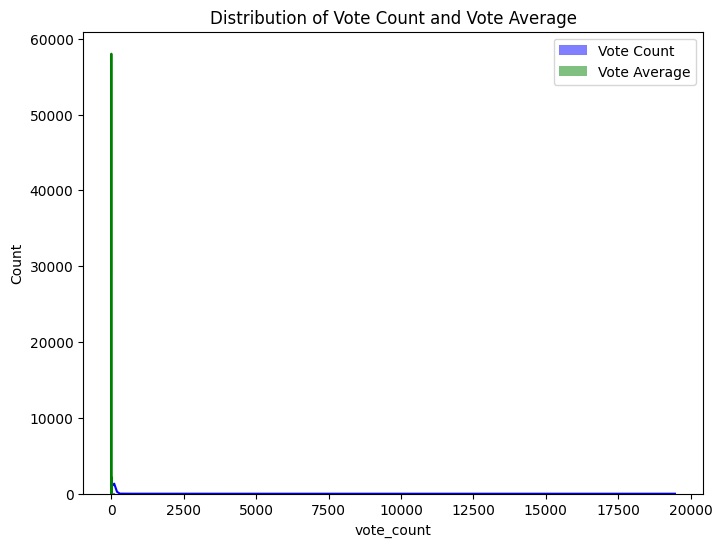
**مصورسازی:**

****

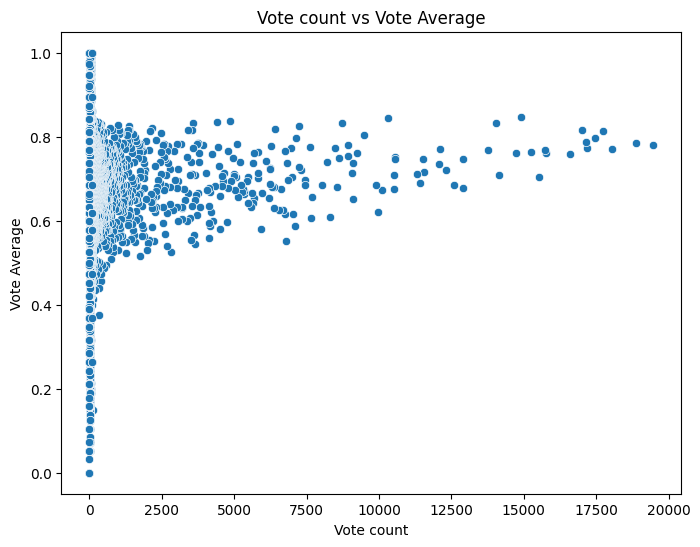
از تحلیل نمودار پراکندگی "بودجه در برابر درآمد" می‌توان نتیجه گرفت که به طور کلی همبستگی مثبتی بین بودجه و درآمد وجود دارد. هرچند بودجه‌های بیشتر معمولاً با درآمدهای بالاتر همراه هستند، اما این تنها عامل تعیین کننده موفقیت نیست. تنوع زیادی در داده‌ها مشاهده می‌شود؛ برخی پروژه‌ها با بودجه کمتر توانسته‌اند درآمدهای بالایی کسب کنند و برخی پروژه‌ها با بودجه بالا درآمدهای نسبتاً کمتری داشته‌اند. این نوسانات نشان می‌دهد که عوامل دیگری همچون کیفیت محتوا، شرایط بازار و زمان‌بندی نیز نقش مهمی ایفا می‌کنند. بنابراین، در برنامه‌ریزی مالی و استراتژی‌گذاری، علاوه بر میزان سرمایه‌گذاری، باید به این عوامل دیگر نیز توجه کرد.

ارتباط بودجه و درآمد:



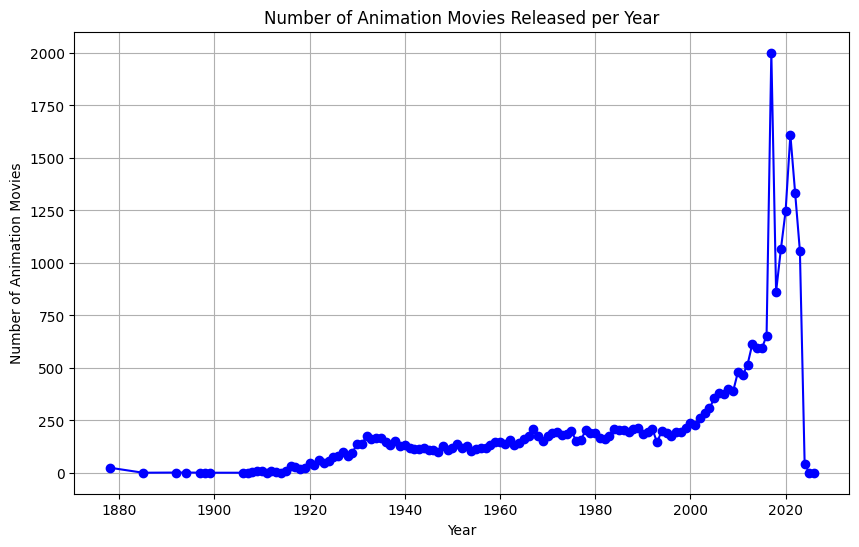
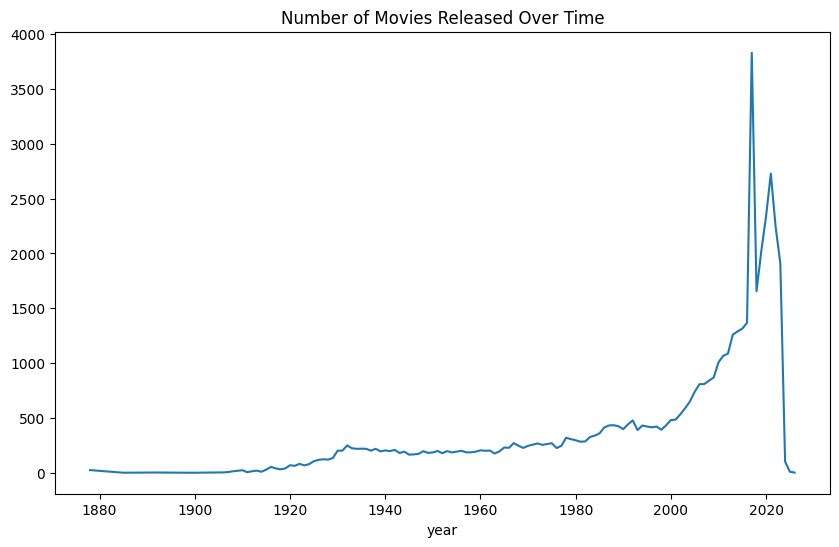


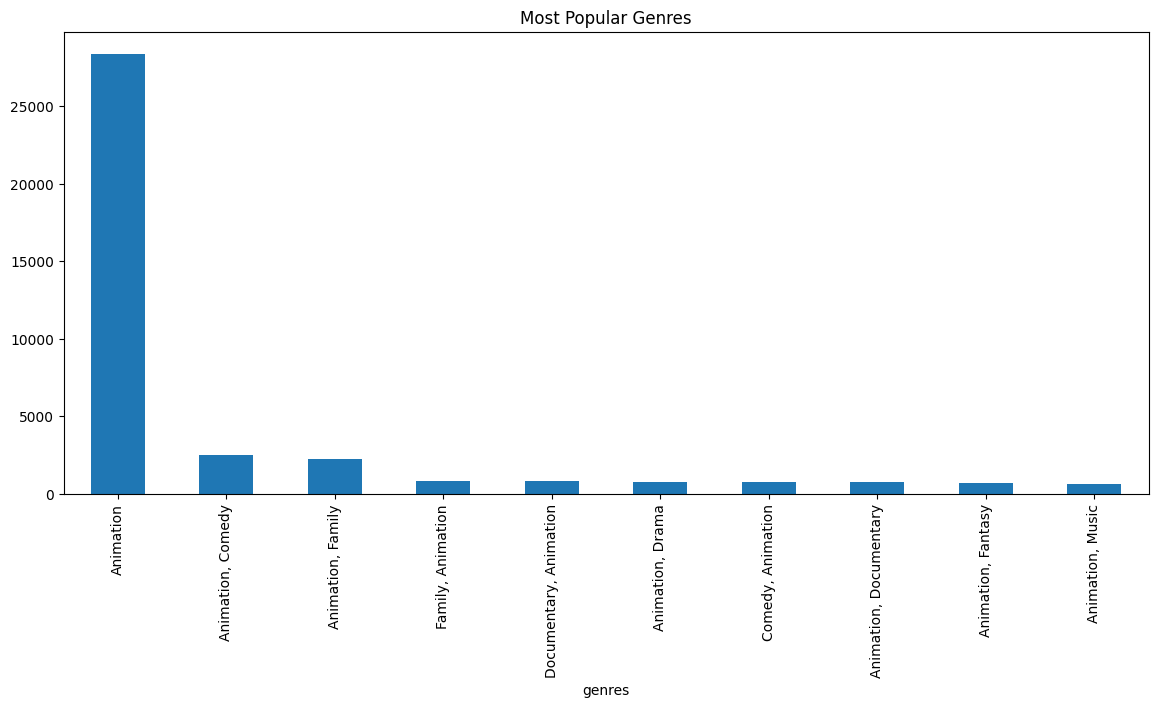
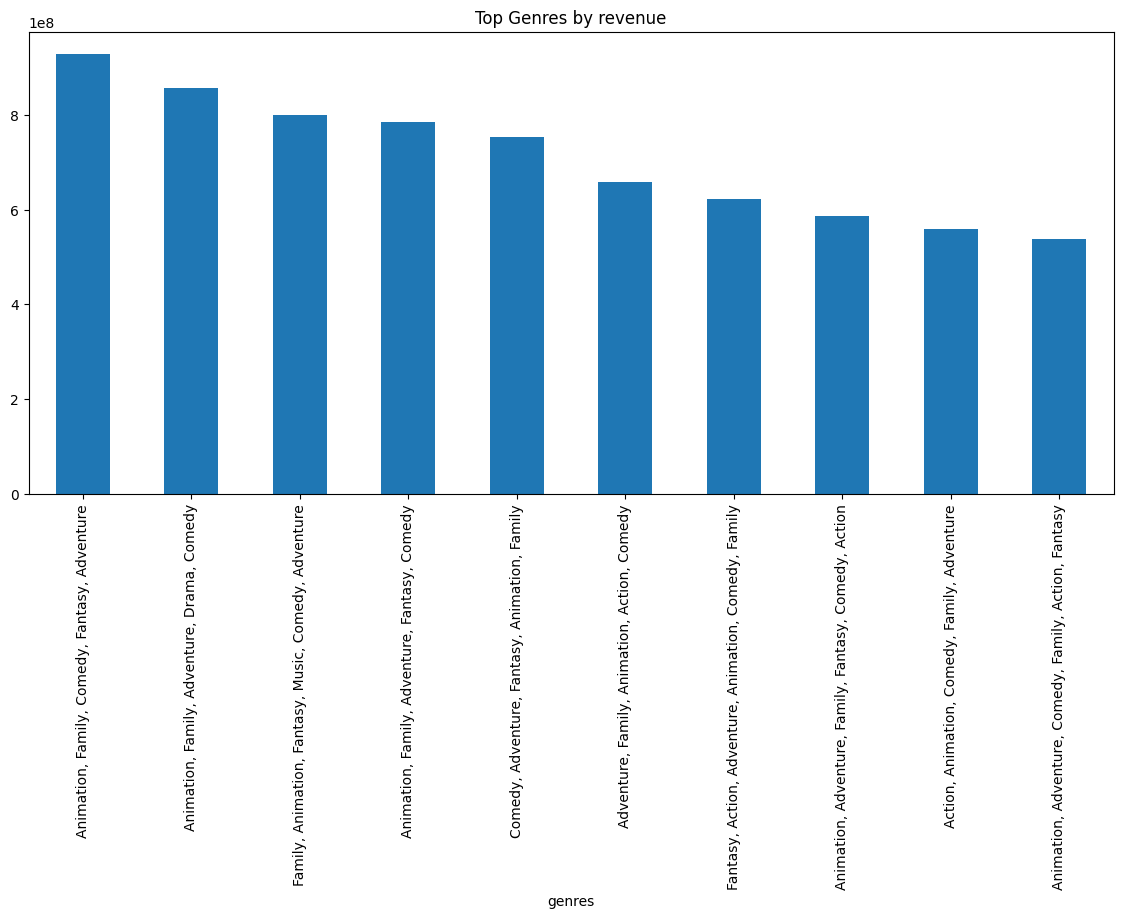
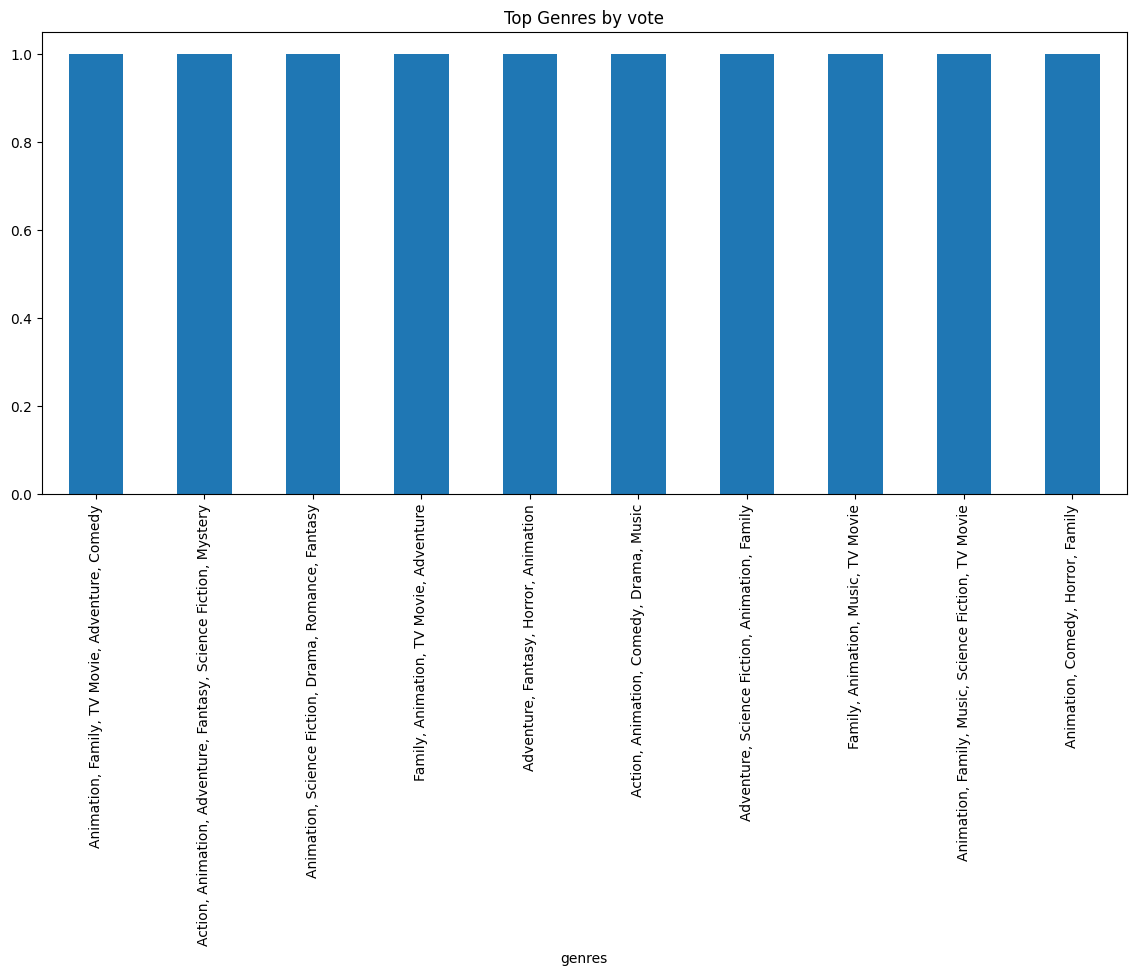
هیستوگرام نشان می‌دهد که اکثر داده‌ها تعداد رای بسیار کمی دارند و یک افزایش قابل توجه در نزدیکی صفر در محور افقی مشاهده می‌شود. این توزیع به سرعت کاهش می‌یابد زیرا تعداد رای‌ها افزایش می‌یابد، که نشان‌دهنده این است که بیشتر آیتم‌ها تعداد رای کمی دارند و تعداد کمی از آیتم‌ها تعداد رای بالایی دارند. این توزیع جالب است زیرا نابرابری در تعداد رای‌ها را برجسته می‌کند و نشان می‌دهد که تعداد کمی از آیتم‌ها بیشتر رای‌ها را دریافت می‌کنند در حالی که بیشتر آیتم‌ها رای‌های بسیار کمی دارند.

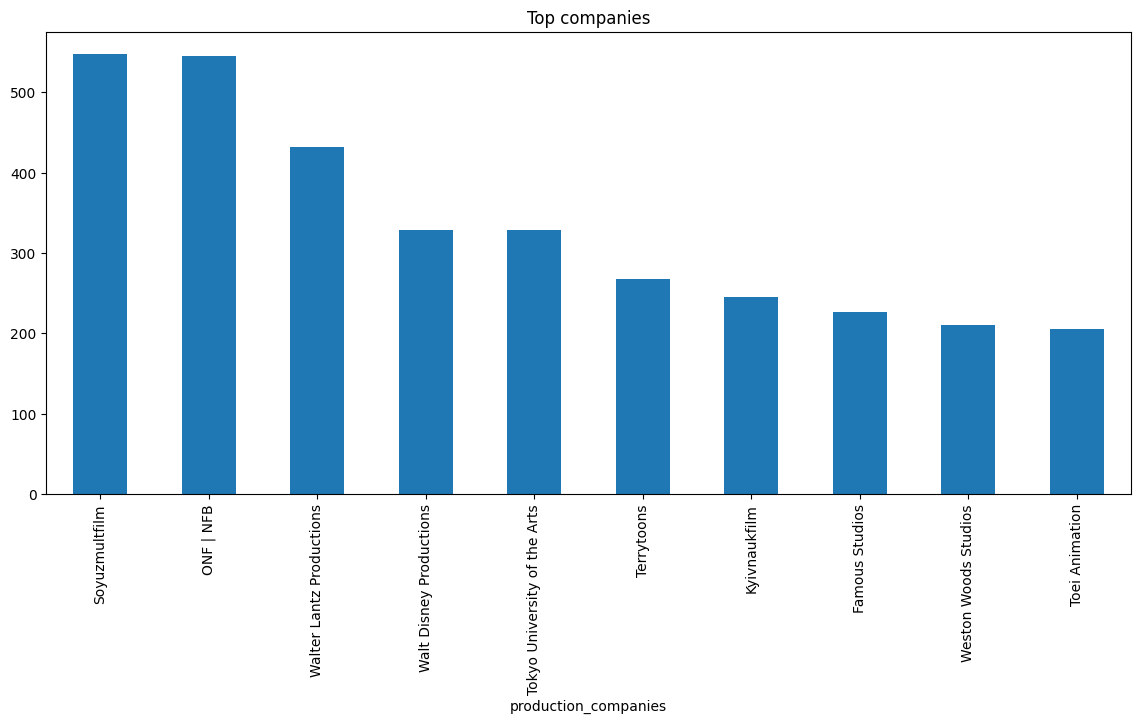


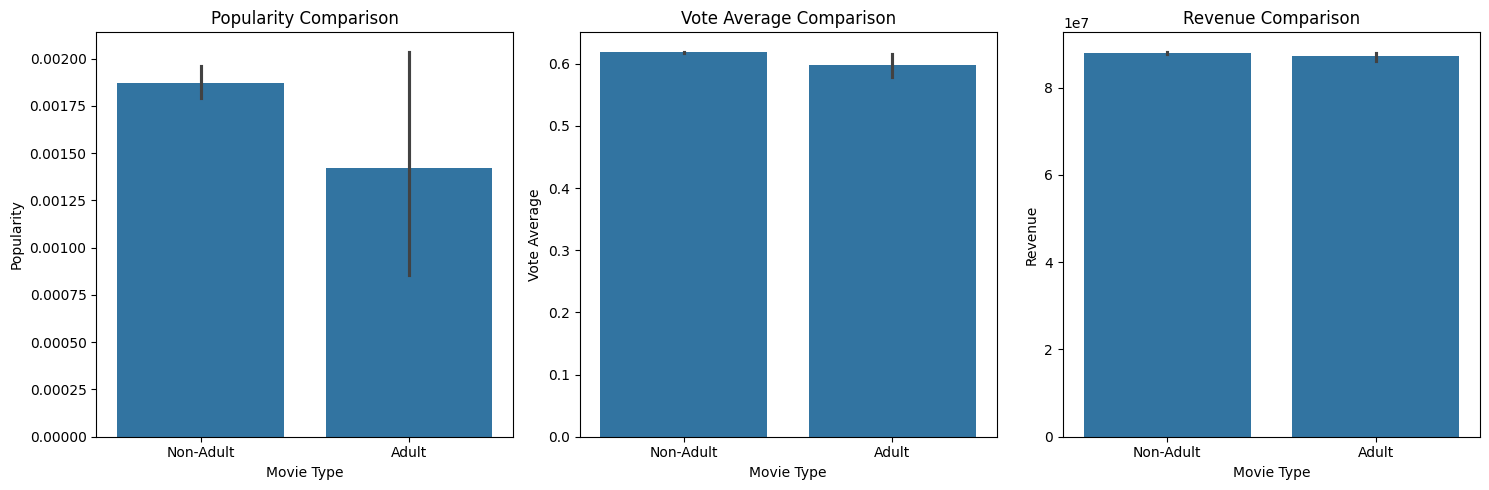
این نمودار نشان می‌دهد که تراکم بالایی از نقاط با تعداد رأی‌های کم (عمدتاً زیر 2500) وجود دارد و میانگین رأی‌ها در این نقاط متغیر است. با افزایش تعداد رأی‌ها، پراکندگی میانگین رأی‌ها کاهش می‌یابد و اکثر نقاط با تعداد رأی‌های بالا میانگین رأی‌های بین 0.6 و 0.9 دارند.

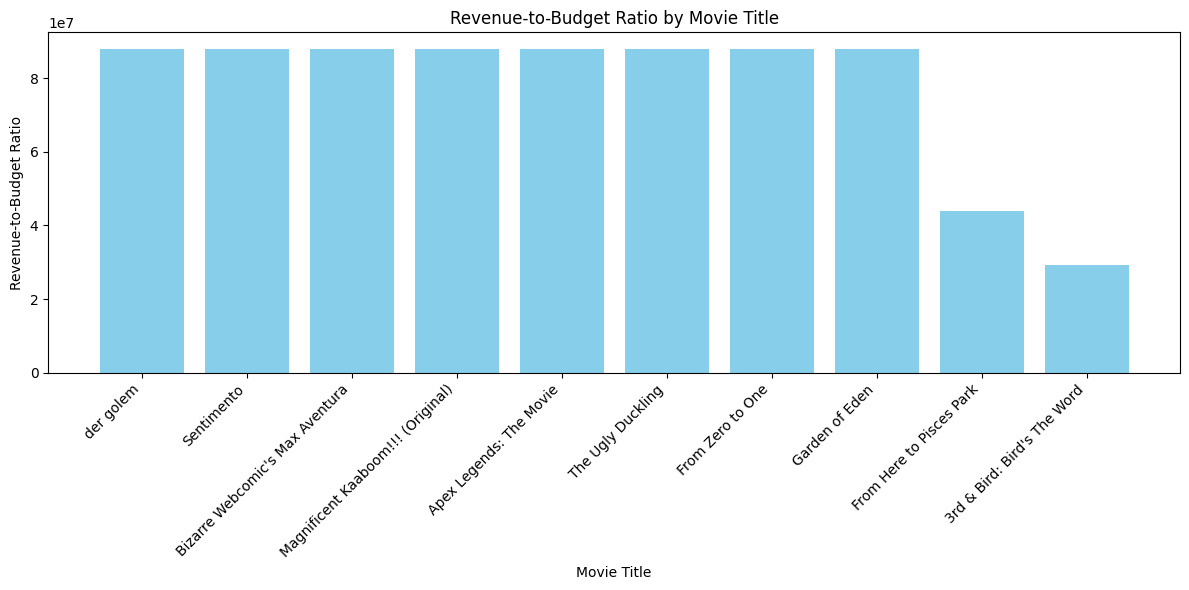
این موضوع نشان می‌دهد که آیتم‌هایی که تعداد رأی بیشتری دارند معمولاً امتیازات بالاتر و پایدارتر دریافت می‌کنند، در حالی که آیتم‌هایی با تعداد رأی کمتر، تنوع بیشتری در امتیازات خود دارند. این الگو ممکن است نشان‌دهنده این باشد که آیتم‌های محبوب‌تر (با تعداد رأی بیشتر) به طور کلی امتیازات مطلوب‌تر و پایدارتر دریافت می‌کنند.

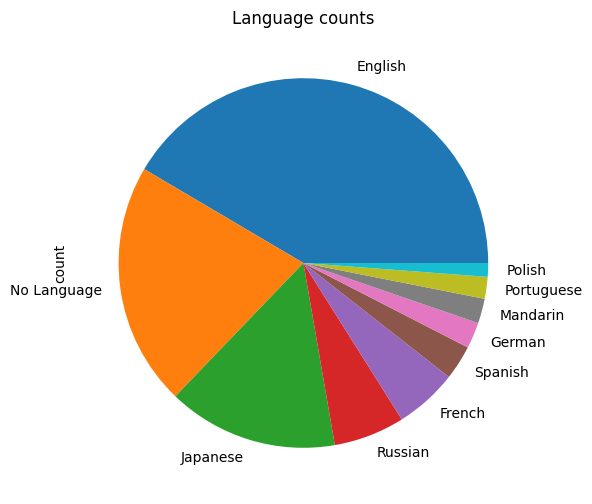
 

انیمیشن محبوب ترین ژانر  









 خلاصه خروجی ها :