فاز دوم پروژه

**توضیح و تحلیل خوشه‌بندی داده‌ها**

**1. خوشه‌بندی بر اساس شباهت ژانر و محتوا (Genre and Content Similarity)**

**ویژگی‌ها:**

* *runtime\_normalized*: مدت زمان فیلم که نرمال‌سازی شده است.
* *content\_complexity*: میزان پیچیدگی محتوای فیلم.
* *genres*: ژانر فیلم‌ها که به صورت کدگذاری دسته‌ای به داده‌های عددی تبدیل شده‌اند.

**تحلیل:**

* فیلم‌هایی با مدت زمان مشابه و پیچیدگی محتوای نزدیک، به همراه شباهت در ژانر، در یک خوشه قرار گرفته‌اند.
* به عنوان مثال، فیلم‌های کمدی با زمان متوسط و محتوای ساده احتمالاً در یک خوشه دسته‌بندی شده‌اند.
* این خوشه‌بندی می‌تواند نشان دهد که فیلم‌های درام یا اکشن با زمان طولانی‌تر و پیچیدگی بیشتر، در خوشه‌های متفاوت قرار دارند.

**2. خوشه‌بندی بر اساس شباهت مالی و محبوبیت (Commercial and Popularity Similarity)**

**ویژگی‌ها:**

* *budget\_normalized*: بودجه فیلم که نرمال‌سازی شده است.
* *revenue\_normalized*: درآمد فیلم که نرمال‌سازی شده است.
* *ROI*: نرخ بازگشت سرمایه.
* *popularity\_normalized*: میزان محبوبیت فیلم.

**تحلیل:**

* فیلم‌هایی با بودجه بالا و ROI پایین احتمالاً در یک خوشه قرار دارند که ممکن است نشان‌دهنده فیلم‌های پرهزینه‌ای باشد که موفقیت تجاری نداشته‌اند.
* فیلم‌های کم‌هزینه با ROI بالا و محبوبیت بالا نیز در خوشه‌های خاص قرار گرفته‌اند که نشان‌دهنده فیلم‌های مستقل موفق است.
* این خوشه‌بندی الگوهایی از موفقیت مالی و محبوبیت فیلم‌ها ارائه می‌دهد.

**3. خوشه‌بندی عمومی (General Combination)**

**ویژگی‌ها:**

* *popularity\_normalized*: محبوبیت نرمال‌شده.
* *vote\_average\_normalized*: میانگین امتیازات مخاطبان که نرمال‌سازی شده است.
* *budget\_normalized*: بودجه نرمال‌شده.
* *runtime\_normalized*: مدت زمان نرمال‌شده.
* *genres*: ژانرها که کدگذاری شده‌اند.

**تحلیل:**

* این خوشه‌بندی ترکیبی از ویژگی‌های مالی، محتوایی و محبوبیت را در نظر می‌گیرد.
* به عنوان مثال، فیلم‌هایی با محبوبیت بالا، امتیاز خوب، و بودجه بالا احتمالاً در یک خوشه قرار دارند.
* این خوشه‌بندی می‌تواند به شناسایی ارتباط میان ویژگی‌های مختلف کمک کند، مانند اینکه آیا محبوبیت با بودجه و مدت زمان رابطه مستقیمی دارد.

**4. خوشه‌بندی زمانی (Temporal Analysis)**

**ویژگی‌ها:**

* *release\_year*: سال انتشار فیلم.
* *release\_season*: فصل انتشار که به صورت کدگذاری دسته‌ای عددی شده است.

**تحلیل:**

* فیلم‌های منتشرشده در یک سال خاص یا فصل خاص، بر اساس ویژگی‌های زمانی در خوشه‌های جداگانه قرار می‌گیرند.
* الگوهایی از تولید فیلم‌ها در فصل‌های خاص، مانند فیلم‌های اکشن در تابستان یا فیلم‌های درام در زمستان، قابل شناسایی است.
* این خوشه‌بندی می‌تواند نشان دهد که آیا فصل انتشار بر موفقیت فیلم تأثیر دارد یا خیر.

**5. خوشه‌بندی زمانی و محتوایی (Temporal and Content Analysis)**

**ویژگی‌ها:**

* *release\_year*: سال انتشار.
* *runtime\_normalized*: مدت زمان نرمال‌شده.
* *release\_season*: فصل انتشار که به صورت کدگذاری عددی شده است.
* *genres*: ژانرها که کدگذاری شده‌اند.

**تحلیل:**

* این خوشه‌بندی ترکیبی از ویژگی‌های زمانی و محتوایی است.
* می‌توان بررسی کرد که آیا ژانرهای خاصی در دوره‌های زمانی یا فصل‌های خاصی محبوب بوده‌اند.
* به عنوان مثال، فیلم‌های اکشن با زمان طولانی در تابستان ممکن است در یک خوشه خاص قرار داشته باشند.

**6. خوشه‌بندی زمانی و مالی (Temporal and Commercial Analysis)**

**ویژگی‌ها:**

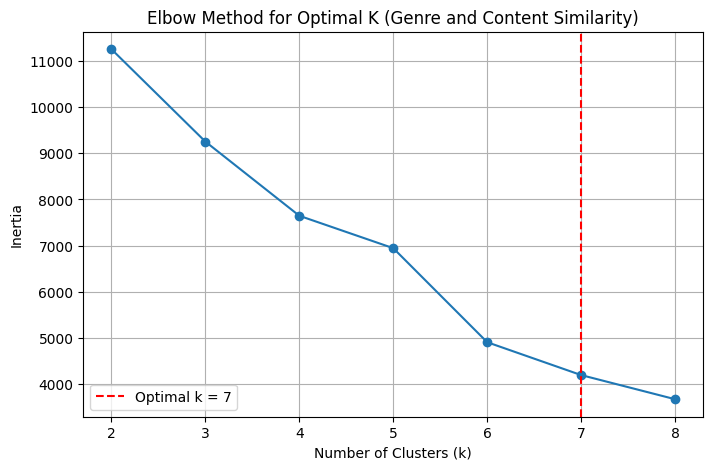
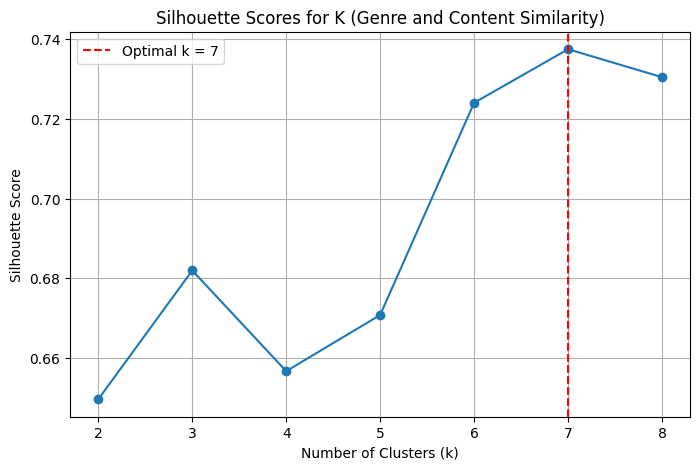
* *release\_year*: سال انتشار.
* *budget\_normalized*: بودجه نرمال‌شده.
* *revenue\_normalized*: درآمد نرمال‌شده.
* *popularity\_normalized*: محبوبیت نرمال‌شده.
* *release\_season*: فصل انتشار که به صورت کدگذاری عددی شده است.

**تحلیل:**

* این خوشه‌بندی ترکیبی از ویژگی‌های زمانی و مالی است.
* می‌توان بررسی کرد که آیا فیلم‌هایی با بودجه بالا در سال یا فصل خاصی منتشر شده‌اند.
* این تحلیل همچنین نشان می‌دهد که فیلم‌های کم‌هزینه با محبوبیت بالا، در کدام دوره‌های زمانی تولید شده‌اند.

**نتیجه کلی:**

این خوشه‌بندی‌ها به ما امکان می‌دهند تا الگوهای معناداری را در داده‌ها شناسایی کنیم و فیلم‌ها را بر اساس شباهت‌های مالی، محتوایی و زمانی در گروه‌های مشابه دسته‌بندی کنیم. این تحلیل می‌تواند در تصمیم‌گیری‌های تجاری و تولید فیلم‌ها بسیار مفید باشد.

**توضیح نمودارها برای شباهت ژانر و محتوا**

**روش Elbow برای تعیین تعداد بهینه خوشه‌ها (Genre and Content Similarity)**

این نمودار نشان‌دهنده **روش Elbow** است که میزان **اینرسی** (جمع مربعات فاصله نقاط از مرکز خوشه‌ها) را برای مقادیر مختلف خوشه‌ها (k) ارزیابی می‌کند. هدف این روش یافتن نقطه‌ای است که کاهش اینرسی به‌طور قابل‌توجهی کند شود (نقطه "آرنج")، زیرا این نقطه تعداد بهینه خوشه‌ها را مشخص می‌کند.

در نمودار ارائه‌شده:

* محور **افقی** تعداد خوشه‌ها (k) را نشان می‌دهد که از 2 تا 8 متغیر است.
* محور **عمودی** مقدار اینرسی را برای هر مقدار k نشان می‌دهد.
* **خط قرمز خط‌چین** تعداد بهینه خوشه‌ها را مشخص می‌کند که **k = 7** است. این نقطه نشان می‌دهد که تقسیم داده‌ها به 7 خوشه ساختار داده‌ها را به‌خوبی نمایش می‌دهد بدون اینکه مدل پیچیده شود.

**امتیاز Silhouette برای تعداد خوشه‌ها (Genre and Content Similarity)**

این نمودار نشان‌دهنده **امتیاز Silhouette** برای مقادیر مختلف k است که میزان شباهت نقاط به خوشه خودشان را نسبت به سایر خوشه‌ها اندازه‌گیری می‌کند. مقدار این امتیاز بین -1 و 1 قرار دارد که:

* امتیاز نزدیک به **1** نشان می‌دهد که نقاط به خوشه خودشان به‌خوبی تعلق دارند و از خوشه‌های دیگر دور هستند.
* امتیاز نزدیک به **0** نشان‌دهنده نقاطی است که در مرز بین خوشه‌ها قرار دارند.
* مقادیر منفی نشان می‌دهند که ممکن است نقاط به‌اشتباه خوشه‌بندی شده باشند.

در نمودار ارائه‌شده:

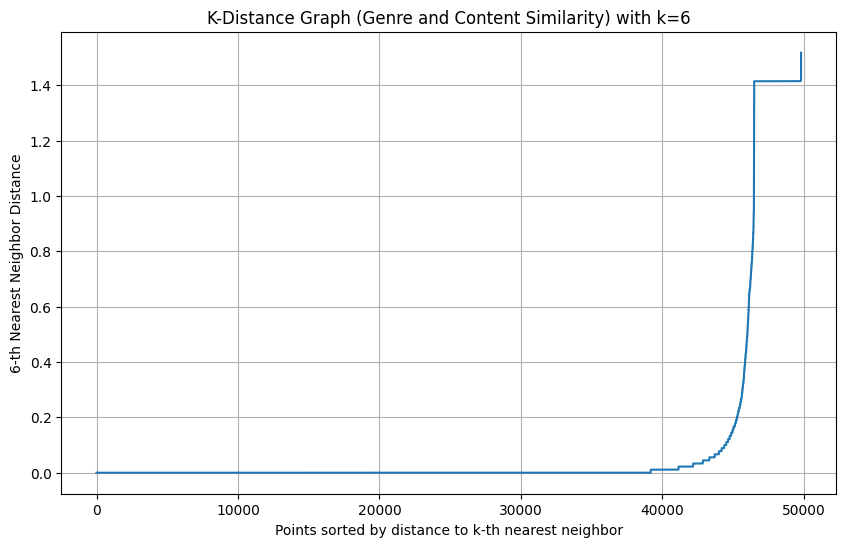
* محور **افقی** تعداد خوشه‌ها (k) را نشان می‌دهد که از 2 تا 8 متغیر است.
* محور **عمودی** امتیاز Silhouette میانگین را برای هر مقدار k نشان می‌دهد.
* **خط قرمز خط‌چین** تعداد بهینه خوشه‌ها را مشخص می‌کند که **k = 7** است. این مقدار بالاترین امتیاز Silhouette (~0.74) را دارد، که نشان‌دهنده حداکثر شباهت درون خوشه‌ای و جداسازی بین خوشه‌ها است.

**تحلیل کلی:**

نمودارهای مربوط به سایر **مجموعه ویژگی‌ها** (مانند شباهت تجاری و محبوبیت، تحلیل زمانی و محتوا) نیز از همین روش پیروی می‌کنند. برای هر مجموعه ویژگی:

1. **روش Elbow** تعداد بهینه k را با مشاهده نمودار اینرسی مشخص می‌کند.
2. **امتیاز Silhouette** انتخاب k را با ارزیابی کیفیت خوشه‌بندی تأیید می‌کند.
3. **k بهینه** بسته به نوع داده‌های هر مجموعه ویژگی متفاوت است.

این تحلیل‌ها تضمین می‌کنند که خوشه‌بندی متناسب با ویژگی‌های خاص هر مجموعه داده طراحی شده و در عین حال قابل تفسیر باقی می‌ماند.



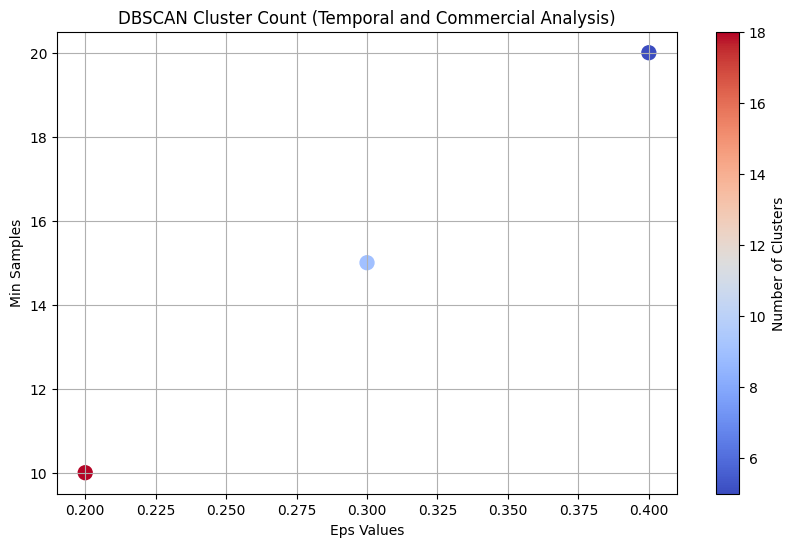
در نمودار **K-Distance Graph** برای فیچر "Genre and Content Similarity" با k=6k = 6، محور افقی نقاط داده را نشان می‌دهد که بر اساس فاصله از ششمین نزدیک‌ترین همسایه مرتب شده‌اند. محور عمودی فاصله این نقاط از ششمین نزدیک‌ترین همسایه‌شان را نمایش می‌دهد.

**تحلیل نمودار:**

* بخش پایینی نمودار، که تقریباً صاف است، نشان‌دهنده نقاطی است که در تراکم بالایی قرار دارند (یعنی خوشه‌ها). این نقاط به یکدیگر نزدیک‌تر هستند.
* نقطه‌ای که شیب نمودار به‌طور قابل‌توجهی افزایش می‌یابد (زانو یا elbow)، معمولاً مرز بین داده‌های خوشه‌ای و نقاط نویز (outliers) را نشان می‌دهد. در این نمودار، این نقطه تقریباً در بازه ۴۰۰۰۰ به ۴۵۰۰۰ مشاهده می‌شود.
* نقاط بالای نمودار، که فاصله زیادی از همسایگان دارند، داده‌هایی هستند که به‌عنوان نویز یا نقاط پرت شناخته می‌شوند.

برای سایر فیچرها نیز روند مشابهی وجود دارد، به این صورت که نمودارها زانویی مشابه خواهند داشت که به کمک آن می‌توان مقدار آستانه مناسب برای DBSCAN را انتخاب کرد.

در نمودار نمایش داده شده برای **"Temporal and Commercial Analysis"**، تعداد خوشه‌ها (Clusters) بر اساس مقادیر eps\text{eps} و min\_samples\text{min\\_samples} در الگوریتم DBSCAN تنظیم شده است.



**توضیحات نمودار:**

1. **محور افقی**: مقدار eps\text{eps} یا شعاع جستجو برای همسایه‌های نزدیک.
2. **محور عمودی**: مقدار min\_samples\text{min\\_samples} که حداقل تعداد نقاط مورد نیاز برای تشکیل یک خوشه است.
3. **رنگ نقاط**: تعداد خوشه‌های تشکیل‌شده برای هر ترکیب از eps\text{eps} و min\_samples\text{min\\_samples}. رنگ‌های گرم‌تر (مانند قرمز) نشان‌دهنده تعداد خوشه‌های بیشتر است و رنگ‌های سردتر (مانند آبی) تعداد کمتری از خوشه‌ها را نشان می‌دهد.
4. **اندازه نقاط**: مربوط به تعداد نقاطی است که نویز (Noise) شناسایی شده‌اند.

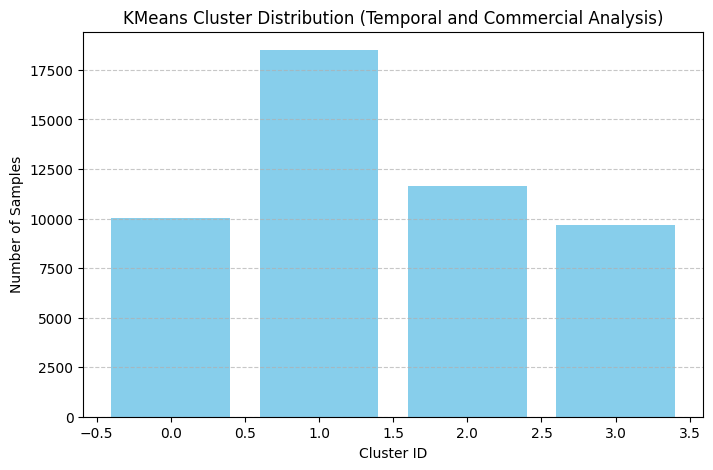
**تحلیل:**

* در مقادیر کوچک eps\text{eps} (مثل 0.2) و min\_samples\text{min\\_samples} کم (10)، تعداد خوشه‌ها زیاد (18) است، اما نقاط نویز نیز بیشتر هستند.
* با افزایش eps\text{eps} به 0.4 و افزایش min\_samples\text{min\\_samples} به 20، تعداد خوشه‌ها به 5 کاهش می‌یابد و نویز بسیار کم می‌شود (6 نقطه)، که نشان‌دهنده خوشه‌بندی بهینه‌تر است.
* مقدار **Silhouette Score** نیز در این حالت به بیشترین مقدار (0.7375) می‌رسد که نشان‌دهنده کیفیت مناسب خوشه‌بندی است.

**نتیجه:**

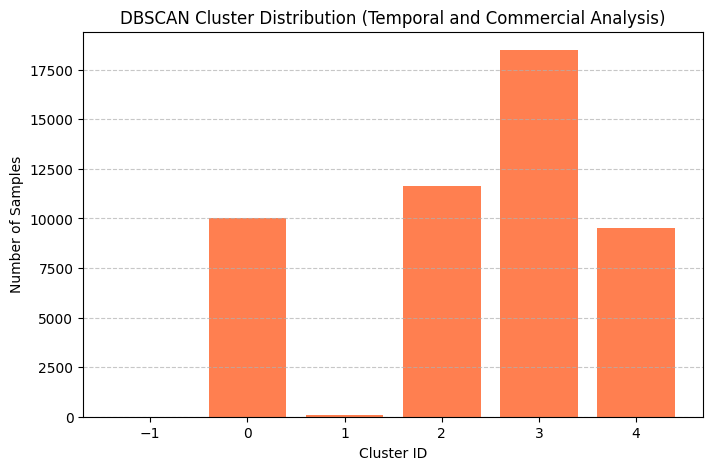
برای فیچر "Temporal and Commercial Analysis"، بهترین تنظیمات DBSCAN با eps=0.4\text{eps} = 0.4 و min\_samples=20\text{min\\_samples} = 20 به‌دست آمده است. نمودارهای سایر فیچرها نیز به همین روش تحلیل می‌شوند و با توجه به تغییر مقادیر eps\text{eps} و min\_samples\text{min\\_samples}، خوشه‌بندی مناسب انتخاب خواهد شد.

**نمودار اول: توزیع خوشه‌های K-Means**



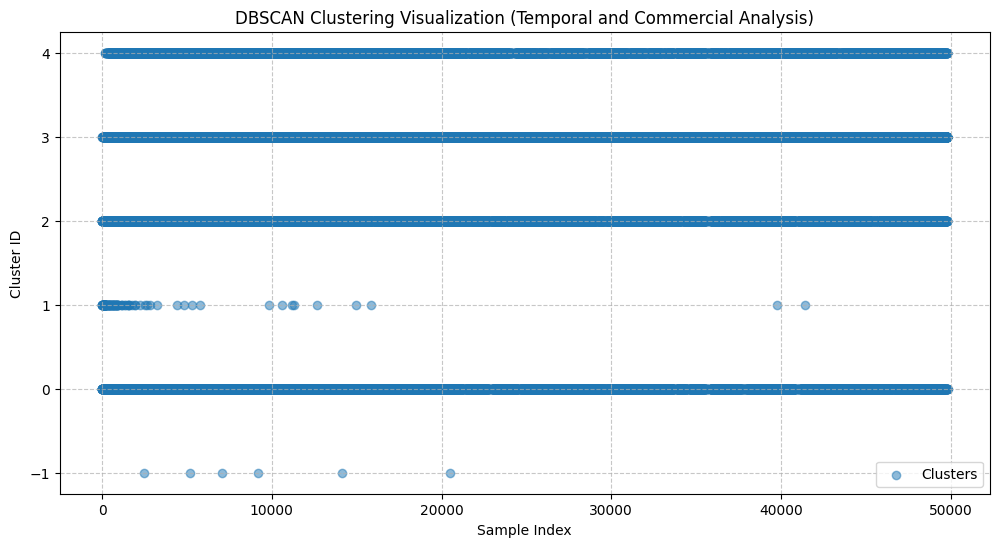
* این نمودار نشان‌دهنده تعداد نمونه‌های اختصاص‌یافته به هر خوشه در الگوریتم K-Means است.
* **مشاهدات:**
  + خوشه شماره 1 دارای بیشترین تعداد نمونه است (حدود 17500).
  + سایر خوشه‌ها تعداد کمتری نمونه دارند که نشان‌دهنده توزیع نابرابر داده‌ها در خوشه‌ها است.
  + K-Means به دلیل فرض اشکال کروی، ممکن است برخی از ساختارهای پیچیده داده را نادیده بگیرد.

**نمودار دوم: توزیع خوشه‌های DBSCAN**



* این نمودار تعداد نمونه‌های خوشه‌بندی‌شده توسط DBSCAN را نشان می‌دهد.
* **مشاهدات:**
  + تعداد نویزها (خوشه -1) بسیار کم است (چند نمونه مشخص در پایین نمودار).
  + خوشه شماره 3 بیشترین نمونه‌ها را شامل می‌شود، مشابه خوشه غالب در K-Means.
  + توزیع خوشه‌ها در DBSCAN بهتر از K-Means با داده‌های پرت سازگار است.

**نمودار سوم: تجسم خوشه‌های DBSCAN**



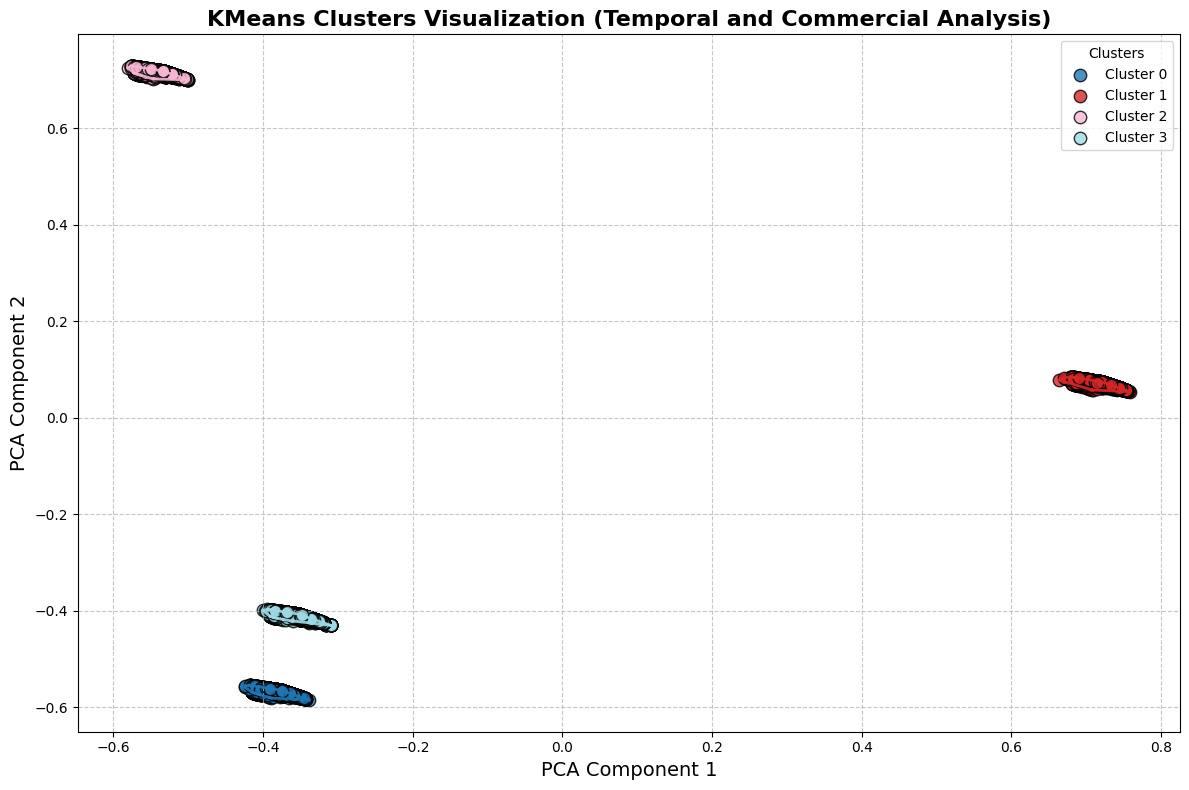
* این نمودار خوشه‌های شناسایی‌شده توسط DBSCAN را در مقابل ایندکس نمونه‌ها نشان می‌دهد.
* **مشاهدات:**
  + نویزها (-1) در انتهای برخی از خوشه‌ها پراکنده‌اند.
  + خوشه‌ها به‌وضوح جدا شده‌اند، که نشان‌دهنده توانایی DBSCAN در تشخیص ساختارهای غیرکروی داده‌ها است.
  + خوشه‌های 0 تا 4 توزیع منظمی در نقاط داده نشان می‌دهند.

**تحلیل کلی:**

* **K-Means** در جداسازی داده‌ها با تعداد بالا بهتر عمل می‌کند اما ممکن است نویزها و خوشه‌های با اشکال پیچیده را نادیده بگیرد.
* **DBSCAN** قادر به مدیریت نویز و خوشه‌های غیرکروی است و خوشه‌بندی انعطاف‌پذیرتری ارائه می‌دهد.
* هر دو روش برای تحلیل "Temporal and Commercial Analysis" مناسب هستند اما DBSCAN برای داده‌های دارای نویز بهتر عمل کرده است.

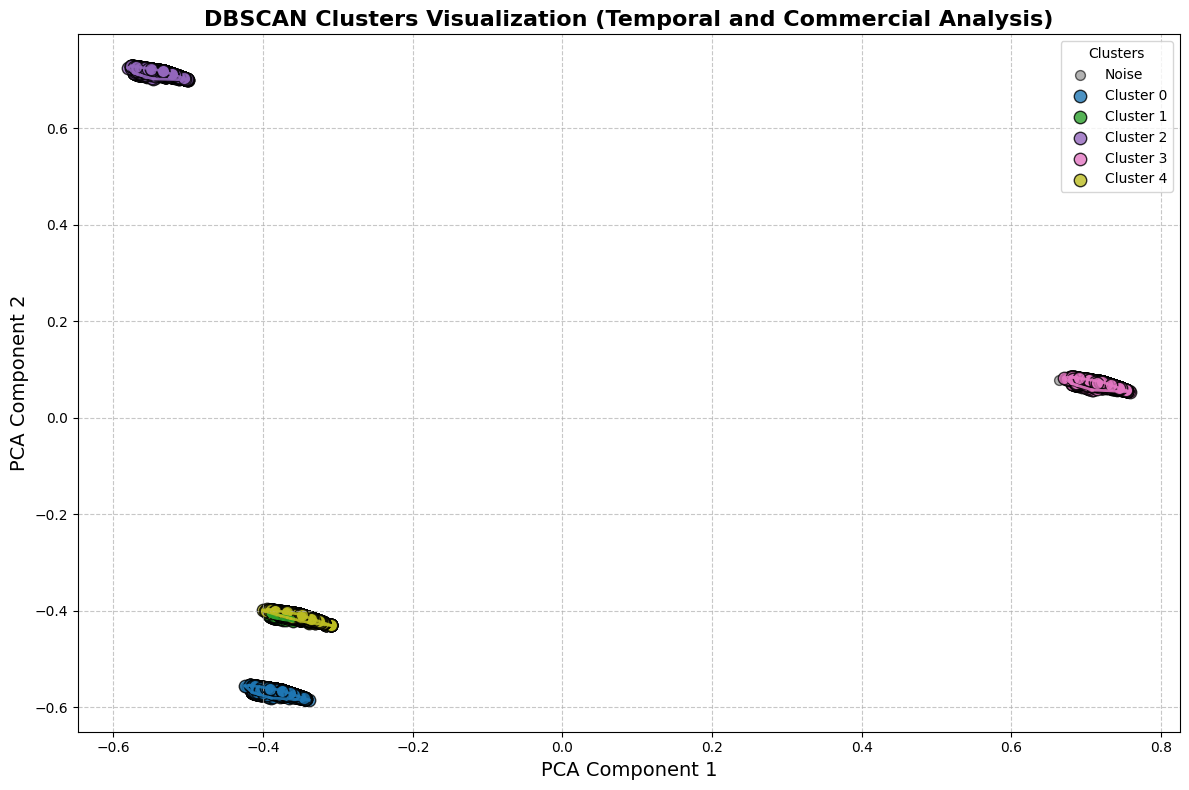
**تحلیل مقایسه‌ای نمودارهای PCA برای K-Means و DBSCAN**

**نمودار اول: KMeans Clusters Visualization**



* این نمودار نمایش خوشه‌بندی K-Means را با استفاده از کاهش ابعاد PCA نشان می‌دهد.
* **مشاهدات:**
  + 4 خوشه (Cluster 0 تا Cluster 3) به‌وضوح قابل‌تشخیص هستند.
  + خوشه‌ها پراکندگی کمی دارند و ساختار کروی فرض شده در K-Means به خوبی رعایت شده است.
  + هیچ داده نویزی مشخص نشده است، چراکه K-Means تمامی نقاط را به خوشه‌ها تخصیص می‌دهد.

**نمودار دوم: DBSCAN Clusters Visualization**



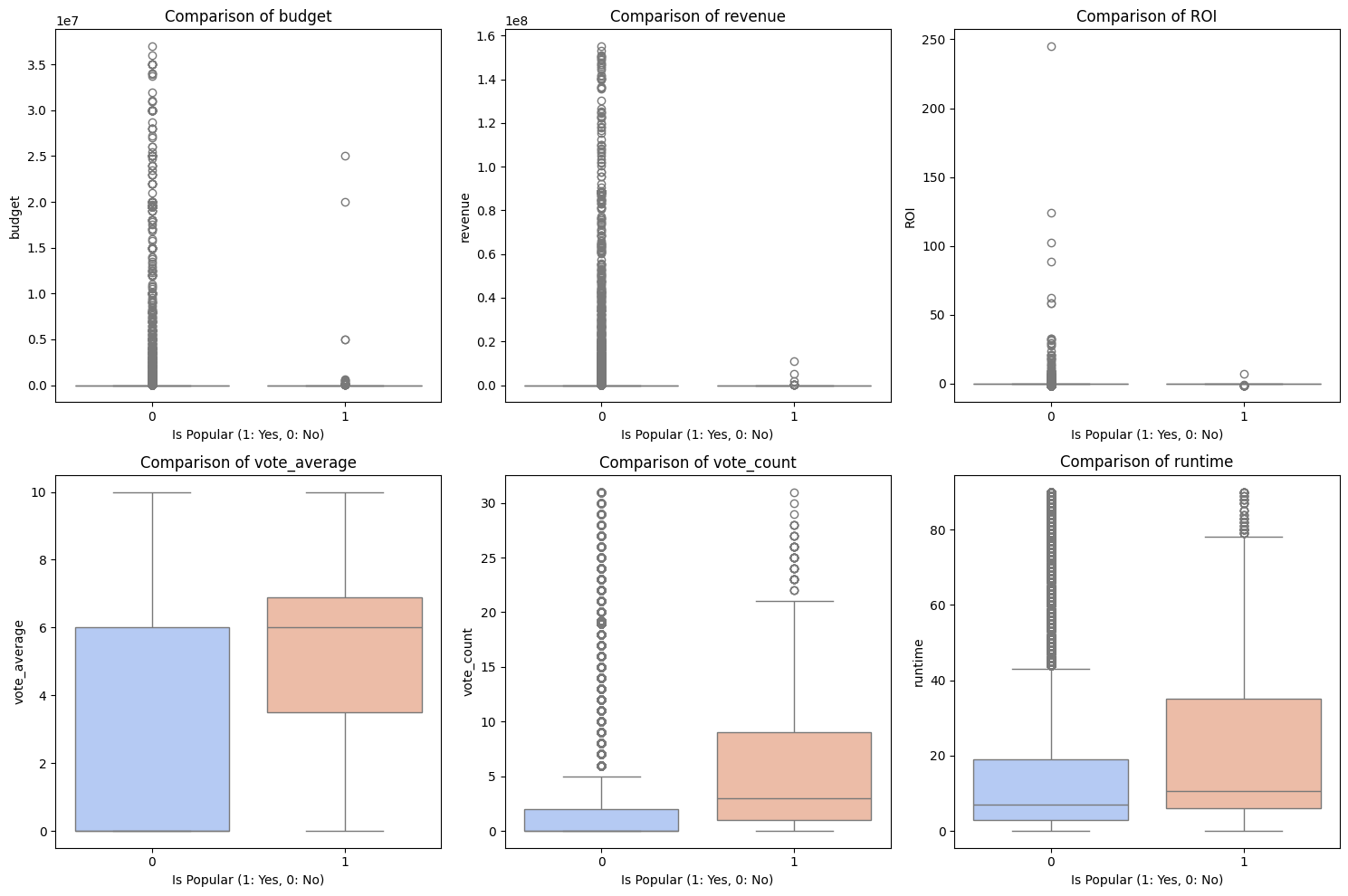
* این نمودار خوشه‌بندی DBSCAN را با استفاده از همان کاهش ابعاد PCA نشان می‌دهد.
* **مشاهدات:**
  + 5 خوشه (Cluster 0 تا Cluster 4) به همراه نقاط نویز (-1) شناسایی شده‌اند.
  + DBSCAN نقاط نویز را که خارج از ساختار خوشه‌ای قرار دارند به درستی مشخص کرده است.
  + خوشه‌ها دارای تراکم مشخص هستند و شکل آن‌ها می‌تواند غیرکروی باشد که DBSCAN با آن سازگار است.

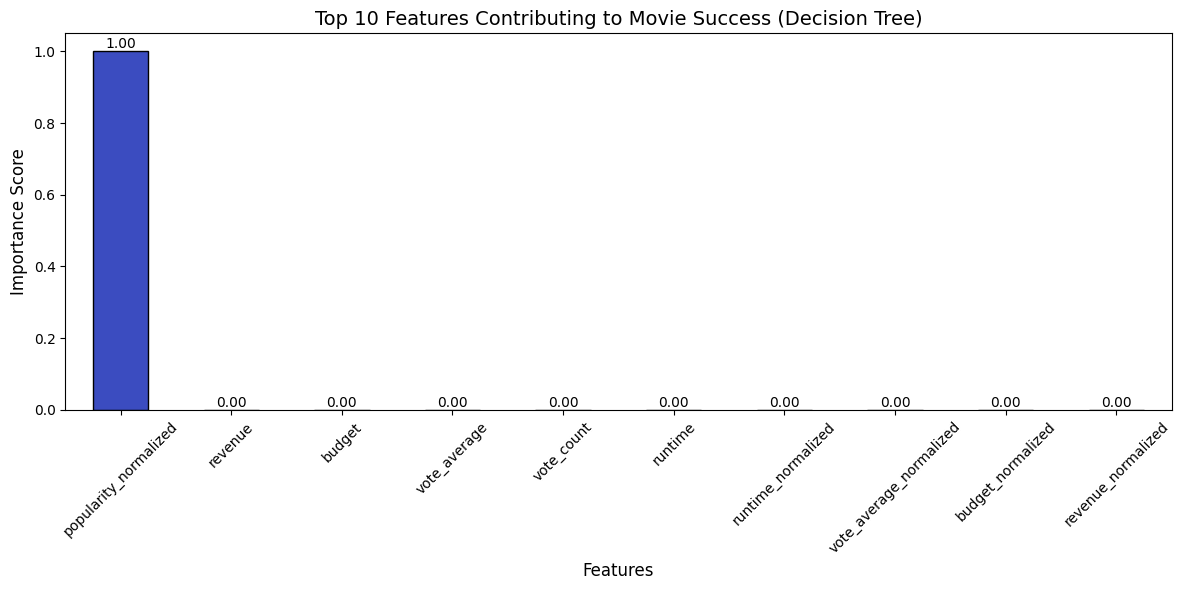
**مقایسه کلی بین K-Means و DBSCAN:**

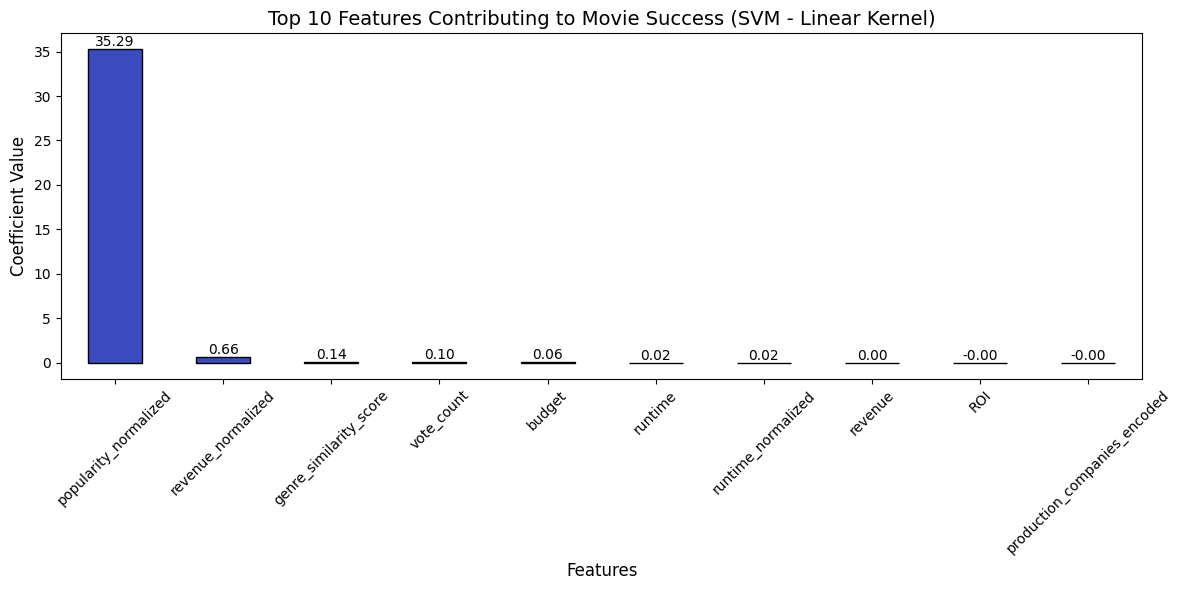
1. **مدیریت نویز:**
   * DBSCAN نویزها را به‌صورت جداگانه مشخص کرده است، درحالی‌که K-Means تمام نقاط را به خوشه‌ها اختصاص داده است.
2. **شکل خوشه‌ها:**
   * K-Means خوشه‌های کروی فرض می‌کند، اما DBSCAN با شکل‌های متنوع‌تری از خوشه‌ها سازگار است.
3. **کاربرد:**
   * برای داده‌هایی که حاوی نویز هستند یا خوشه‌های غیرکروی دارند (مانند داده‌های پیچیده)، DBSCAN مناسب‌تر است.
   * K-Means برای داده‌های با توزیع ساده و کروی مناسب‌تر است.

**نتیجه‌گیری:**

* در تحلیل "Temporal and Commercial Analysis"، استفاده از DBSCAN برای شناسایی نویز و انعطاف بیشتر در خوشه‌بندی مناسب‌تر است.
* K-Means به دلیل محدودیت در فرض کروی بودن خوشه‌ها ممکن است برای این داده‌ها تمامی اطلاعات ساختاری را منعکس نکند.
* در سایر فیچرست‌ها نیز روند خوشه‌بندی مشابه است. K-Means خوشه‌ها را با فرض ساختار کروی و بدون شناسایی نویز تقسیم‌بندی می‌کند، در حالی که DBSCAN خوشه‌های متراکم‌تر را شناسایی کرده و نقاط نویز را جدا می‌کند. ویژگی‌های اصلی هر فیچرست مانند شکل خوشه‌ها و وجود نویز، به طور مشابه در این دو روش تحلیل می‌شوند و رفتار مشابهی با نمودارهای ارائه‌شده دارند.

**دسته بندى(Classification)**





**تحلیل نمودارها و نتایج مدل‌ها**

**تحلیل نمودارها**

1. **نمودار مقایسه ویژگی‌ها بر اساس محبوبیت:**
   * **بودجه (Budget):** فیلم‌های محبوب و غیرمحبوب بودجه‌های مشابهی در اکثر موارد دارند، اما برخی فیلم‌های محبوب بودجه‌های بسیار بالاتری دارند.
   * **درآمد (Revenue):** فیلم‌های محبوب درآمد بیشتری نسبت به فیلم‌های غیرمحبوب داشته‌اند.
   * **بازگشت سرمایه (ROI):** مقادیر بالا در برخی فیلم‌های غیرمحبوب نشان می‌دهد که فیلم‌های با بودجه کم ممکن است بازدهی بالایی داشته باشند.
   * **میانگین رأی‌ها (Vote Average):** فیلم‌های محبوب میانگین رأی بیشتری دارند، که کیفیت را نشان می‌دهد.
   * **تعداد رأی‌ها (Vote Count):** فیلم‌های محبوب تعداد رأی‌های بسیار بیشتری دریافت کرده‌اند.
   * **مدت زمان (Runtime):** فیلم‌های محبوب مدت زمان طولانی‌تری داشته‌اند.
2. **نمودار اهمیت ویژگی‌ها در مدل‌های Decision Tree و SVM:**
   * **Decision Tree:** ویژگی popularity\_normalized تنها ویژگی مهم شناخته شده است.
   * **SVM:** ویژگی popularity\_normalized همچنان بیشترین تأثیر را دارد، اما ویژگی‌های دیگر مانند revenue\_normalized و genre\_similarity\_score نیز تأثیر کمی دارند.

**تحلیل نتایج مدل‌ها**

**۱. Decision Tree:**

* **دقت (Accuracy):** ۱.۰۰۰۰
* **Precision، Recall، F1-Score:** ۱.۰۰۰۰
* **نتیجه:** مدل به طور کامل داده‌های تست را شناسایی کرده و هیچ خطایی نداشته است. این ممکن است ناشی از **Overfitting** باشد.

**۲. SVM:**

* **دقت (Accuracy):** ۰.۹۹۹۱
* **Precision:** ۰.۹۶۳۴
* **Recall:** ۱.۰۰۰۰
* **F1-Score:** ۰.۹۸۱۴
* **نتیجه:** عملکرد بسیار خوبی روی داده‌های تست دارد و از Overfitting جلوگیری کرده است. این مدل تعادل بهتری بین دقت و فراخوانی برقرار کرده است.

**۳. Naive Bayes:**

* **دقت (Accuracy):** ۰.۰۴۳۳
* **Precision:** ۰.۰۲۴۲
* **Recall:** ۰.۹۹۵۸
* **F1-Score:** ۰.۰۴۷۲
* **نتیجه:** این مدل عملکرد ضعیفی داشته و نمی‌تواند فیلم‌های محبوب را به درستی پیش‌بینی کند. این ضعف ممکن است به دلیل فرضیات ساده Naive Bayes در مورد استقلال ویژگی‌ها باشد.

**نتیجه‌گیری کلی:**

* **مدل Decision Tree:** اگرچه دقت بالایی دارد، اما احتمال Overfitting زیاد است و نمی‌تواند تعمیم خوبی روی داده‌های جدید داشته باشد.
* **مدل SVM:** بهترین تعادل بین دقت و قابلیت تعمیم را دارد و برای این داده‌ها مناسب‌ترین مدل است.
* **مدل Naive Bayes:** به دلیل فرضیات ساده خود نمی‌تواند با پیچیدگی داده‌ها به خوبی کنار بیاید.

**پیشنهادات:**

* از **SVM** به عنوان مدل اصلی استفاده کنید.
* برای جلوگیری از Overfitting در Decision Tree از روش‌هایی مانند **Pruning** استفاده کنید.
* ویژگی‌های بیشتری را برای افزایش دقت Naive Bayes بررسی کنید یا مدل را با داده‌های ساده‌تر آزمایش کنید.

**تحلیل آماری ویژگی‌های کلیدی برای فیلم‌های موفق و ناموفق**

**بینش‌های اصلی از داده‌ها:**

1. **بودجه (Budget):**
   * میانگین بودجه فیلم‌های موفق حدود **۴۹ هزار دلار** است، که بسیار بیشتر از میانگین بودجه فیلم‌های ناموفق (۲۰۵ هزار دلار) است. این اختلاف نشان می‌دهد که بودجه بالاتر با موفقیت ارتباط دارد.
2. **درآمد (Revenue):**
   * فیلم‌های موفق به‌طور میانگین **۱.۵۹ میلیون دلار** درآمد دارند، در حالی که فیلم‌های ناموفق به‌طور میانگین تنها **۷۰ هزار دلار** درآمد دارند. این نشان‌دهنده پذیرش بهتر فیلم‌های موفق در بازار است.
3. **بازگشت سرمایه (ROI):**
   * میانگین ROI برای فیلم‌های موفق **۰.۰۳** است، که نشان می‌دهد این فیلم‌ها بودجه خود را بهینه استفاده کرده‌اند. فیلم‌های ناموفق به‌طور میانگین ROI منفی دارند، که بیانگر استفاده ناکارآمد از بودجه است.
4. **میانگین رأی‌ها (Vote Average):**
   * میانگین امتیاز فیلم‌های موفق **۴.۹۲** است، در حالی که فیلم‌های ناموفق تنها **۲.۶۳** دارند. این نشان‌دهنده رضایت بیشتر تماشاگران از فیلم‌های موفق است.
5. **تعداد رأی‌ها (Vote Count):**
   * فیلم‌های موفق میانگین **۵.۷** رأی دارند، که تقریباً دو برابر تعداد رأی‌های فیلم‌های ناموفق (۳.۰۹) است. این نشان می‌دهد که فیلم‌های موفق توجه بیشتری را جلب کرده‌اند.
6. **مدت زمان (Runtime):**
   * فیلم‌های موفق میانگین مدت زمان **۲۴ دقیقه** دارند، که بیشتر از فیلم‌های ناموفق با میانگین **۱۶ دقیقه** است. این نشان می‌دهد که یک مدت زمان مناسب می‌تواند برای موفقیت فیلم مؤثر باشد.

**دلایل احتمالی شکست فیلم‌ها:**

1. **بودجه پایین:** بودجه محدود ممکن است منجر به کاهش کیفیت تولید و تبلیغات شود.
2. **ROI منفی:** مدیریت نادرست بودجه و تبلیغات ممکن است منجر به بازدهی ضعیف شود.
3. **رأی‌ها و امتیازهای پایین:** عدم رضایت مخاطبان و نقدهای منفی می‌تواند منجر به شکست شود.
4. **مدت زمان نامناسب:** فیلم‌هایی با مدت زمان بسیار کوتاه یا بیش از حد طولانی ممکن است توجه مخاطبان را از دست بدهند.
5. **ژانر غیرمحبوب:** فیلم‌هایی که ژانرهای کم‌طرفدار یا اشباع‌شده را هدف قرار می‌دهند ممکن است موفقیت کمتری داشته باشند.
6. **شرکت‌های تولیدکننده غیرمعتبر:** عدم همکاری با شرکت‌های تولیدی معتبر ممکن است بر توزیع و بازاریابی فیلم تأثیر منفی بگذارد.

**عملکرد مدل‌ها**

1. **Naive Bayes:**
   * عملکرد بسیار ضعیف، که نشان می‌دهد این مدل برای داده‌های پیچیده مناسب نیست و نیاز به انتخاب ویژگی‌های بهتر یا متعادل‌تر شدن داده‌ها دارد.
2. **Decision Tree:**
   * با دقت ۱۰۰٪ نشان‌دهنده Overfitting است و ممکن است روی داده‌های جدید عملکرد ضعیفی داشته باشد.
3. **SVM:**
   * بهترین تعادل بین دقت و تعمیم‌پذیری را ارائه می‌دهد. این مدل می‌تواند تفاوت‌های خطی بین فیلم‌های موفق و ناموفق را به خوبی تشخیص دهد.

**پیشنهادات:**

* **افزایش بودجه:** تلاش برای تأمین بودجه بیشتر جهت بهبود کیفیت تولید.
* **تمرکز بر ژانرهای محبوب:** تحلیل ژانرها و هدف قرار دادن مخاطبان اصلی.
* **بهینه‌سازی مدت زمان:** ایجاد تعادل در زمان فیلم برای حفظ توجه مخاطبان.
* **همکاری با شرکت‌های معتبر:** استفاده از توانایی‌های بازاریابی و توزیع شرکت‌های معتبر.
* **استفاده از SVM:** این مدل برای پیش‌بینی موفقیت فیلم‌ها بهترین گزینه است و می‌تواند در کنار Decision Tree برای تحلیل‌های عمیق‌تر استفاده شود.