به نام خدا

گزارش پروژه - علم داده **نیایش خانی ۹۹۵۲۱۲۳۵**

هدف این پروژه پیش‌بینی درآمد باکس آفیس فیلم‌ها بر اساس عوامل مختلفی مانند ژانر، بازیگران، کارگردانان، بودجه تولید و ... است. برای رسیدن به این هدف از مجموعه داده‌های Rotten Tomatoes استفاده کرده‌ایم. مراحل انجام این کار طبق موارد خواسته شده در داکیومنت سوال شامل تحلیل داده های اکتشافی (EDA)، پیش پردازش‌داده ها، مهندسی ویژگی، آموزش مدل و ارزیابی است.

نگاهی اولیه به دیتا

بعد از انجام ست‌آپ های اولیه مانند ایمپورت کردن کتابخانه‌ها و لود کردن دیتا، نگاهی اجمالی به دیتا می‌اندازیم و با فیچرها، تایپ داده ها، مقادیر آنها و ... آشنا می‌شویم. فانکشن‌هایی مانند ()info. ، ()head و ()describe به ما در انجام این کار کمک می‌کنند. (به دلیل استفاده مکرر از این موارد در تمرین‌ها از توضیح آنها میگذریم.)

Preprocessing & Data Cleaning

Handling Missing Values

برای پاکسازی دیتا و اطمینان از درستی آن برای انجام تحلیل و آنالیز ، باید اقداماتی انجام داد. اولین کار چک کردن مقادیر null یا Nan است که با دستور ()isnull().sum می‌توان این کار را انجام داد. پس از بدست آوردن اطلاعات از این مقادیر null، یا باید این رکوردها را حذف (drop) و یا با مقادیر مناسب جایگزین کرد. از آنجا که در راهکار اول امکان از دست رفتن داده‌های مهم است، من راه دوم را انتخاب کردم. همانطور که مشخص است فیچرهای rt_website و rt_tagline مقادیر نال دارند که آنها را با Unknown پر کردم. همچنین داده دیگری که می‌تواند مشکل ساز باشد rt_release_date است که timestamp است و من آن را بوسیله to_datetime هندل کردم.

Parsing JSON Columns

چندین ستون دیگر نیز از جمله rt_genres، rt_keywords، rt_languages، و rt_production_countries در دیتافریم movies و ستون‌های rt_actors و rt_staff در دیتای credit دردسرساز بودند چراکه شامل رشته‌های JSON مانند بودند. بنابراین باید آنها را تجزیه می‌کردیم. بوسیله این فانکشن این ستون‌ها به لیست‌های پایتون تبدیل شدند که به ما این امکان را می‌دهند تا آن‌ها را بیشتر تحلیل کنیم.

تجزیه و تحلیل داده های اکتشافی (EDA)

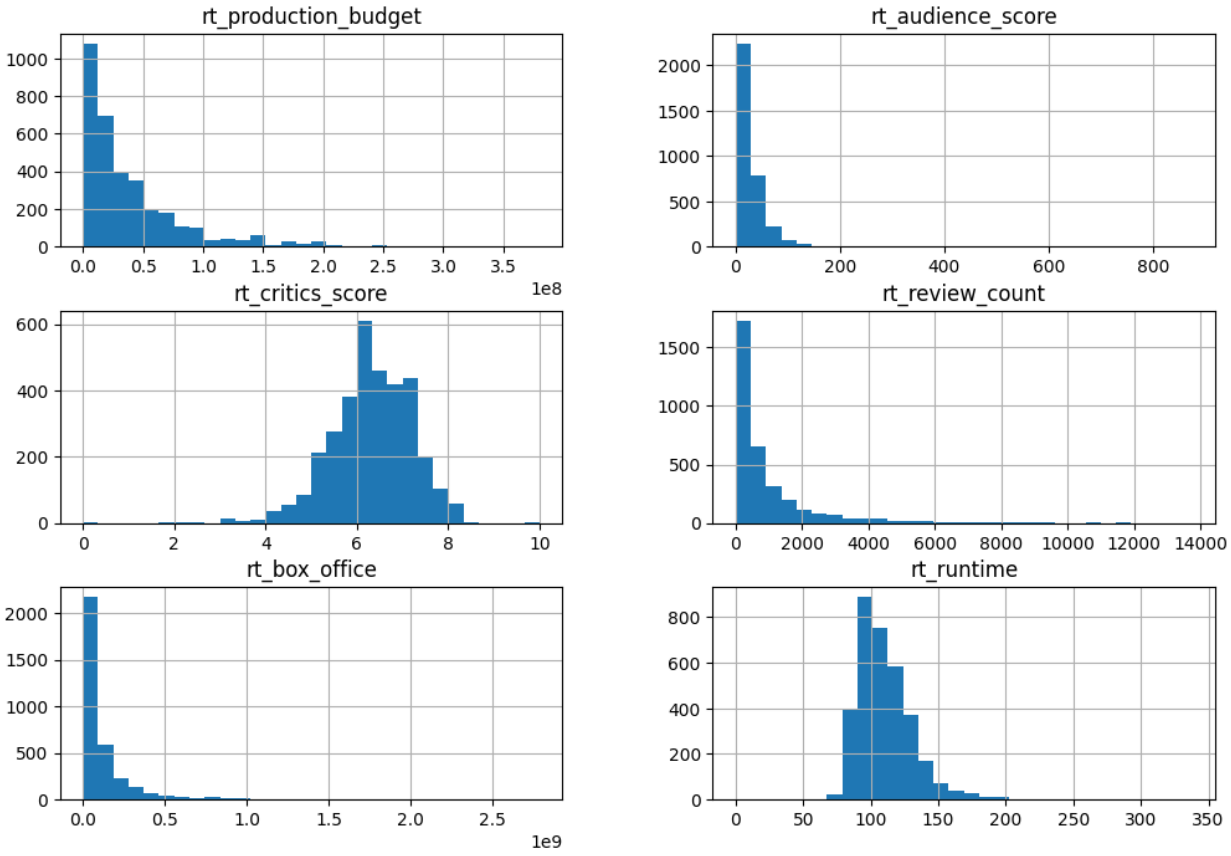
در بخش EDA، برای درک بهتر مجموعه داده، شناسایی الگوها و استخراج بینش مفید، بر پاسخ به سوالات مطرح شده تمرکز کردیم. حال به بررسی این سوالات و روش های مورد استفاده برای پاسخ به آنها می‌پردازیم:

کشیدن هیستوگرام برای داده‌های عددی:

```
# Select numerical columns
num_cols = ['rt_production_budget', 'rt_audience_score', 'rt_critics_score', 'rt_review_count',
'rt_box_office', 'rt_runtime']

# Plot histograms
df_movies[num_cols].hist(figsize=(12, 8), bins=30)
plt.show()
```

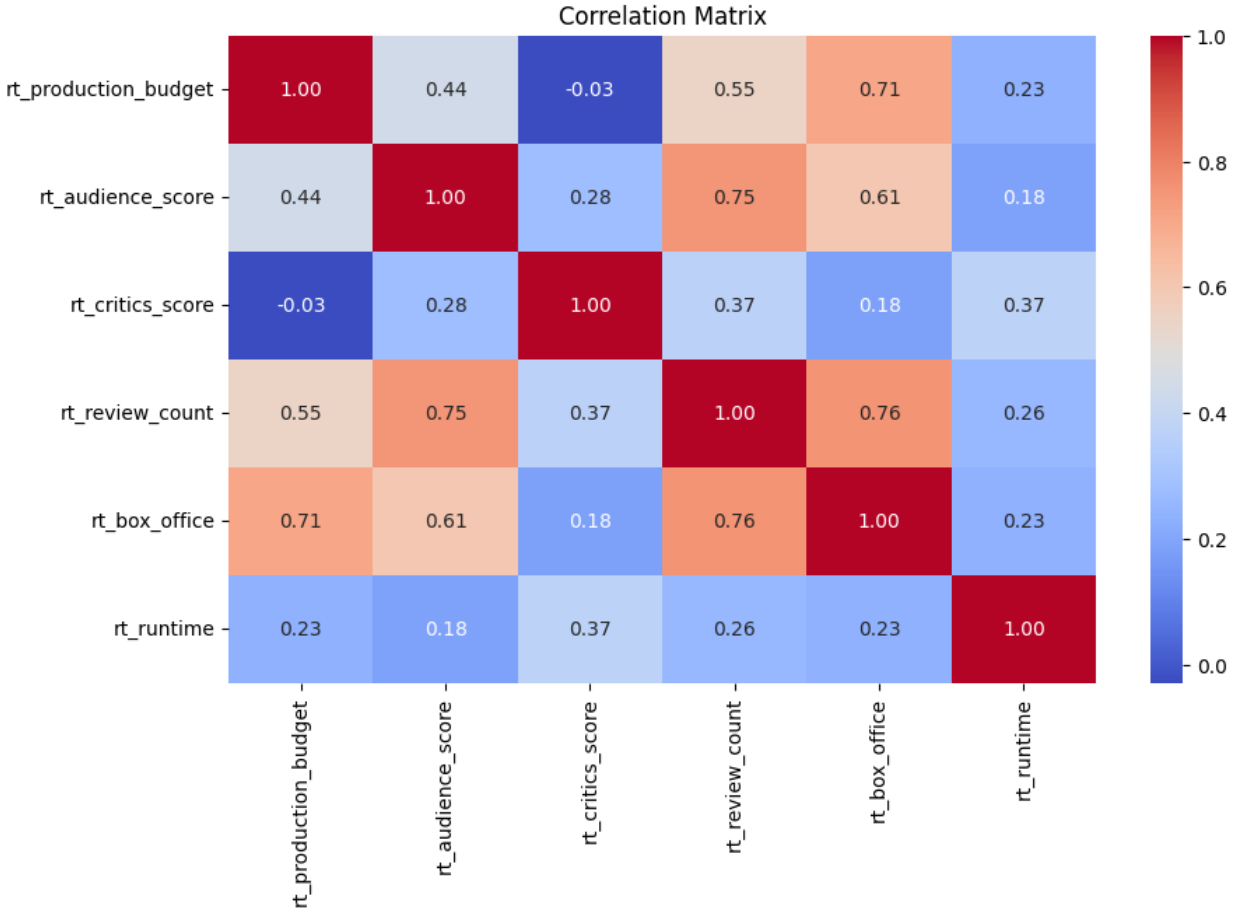
خروجی:



بررسی correlation و میزان همبستگی میان فیچرها و rt_box_office که تارگت ما است:

```
# Correlation heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(df_movies[num_cols].corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Matrix")
plt.show()
```

خروجی:



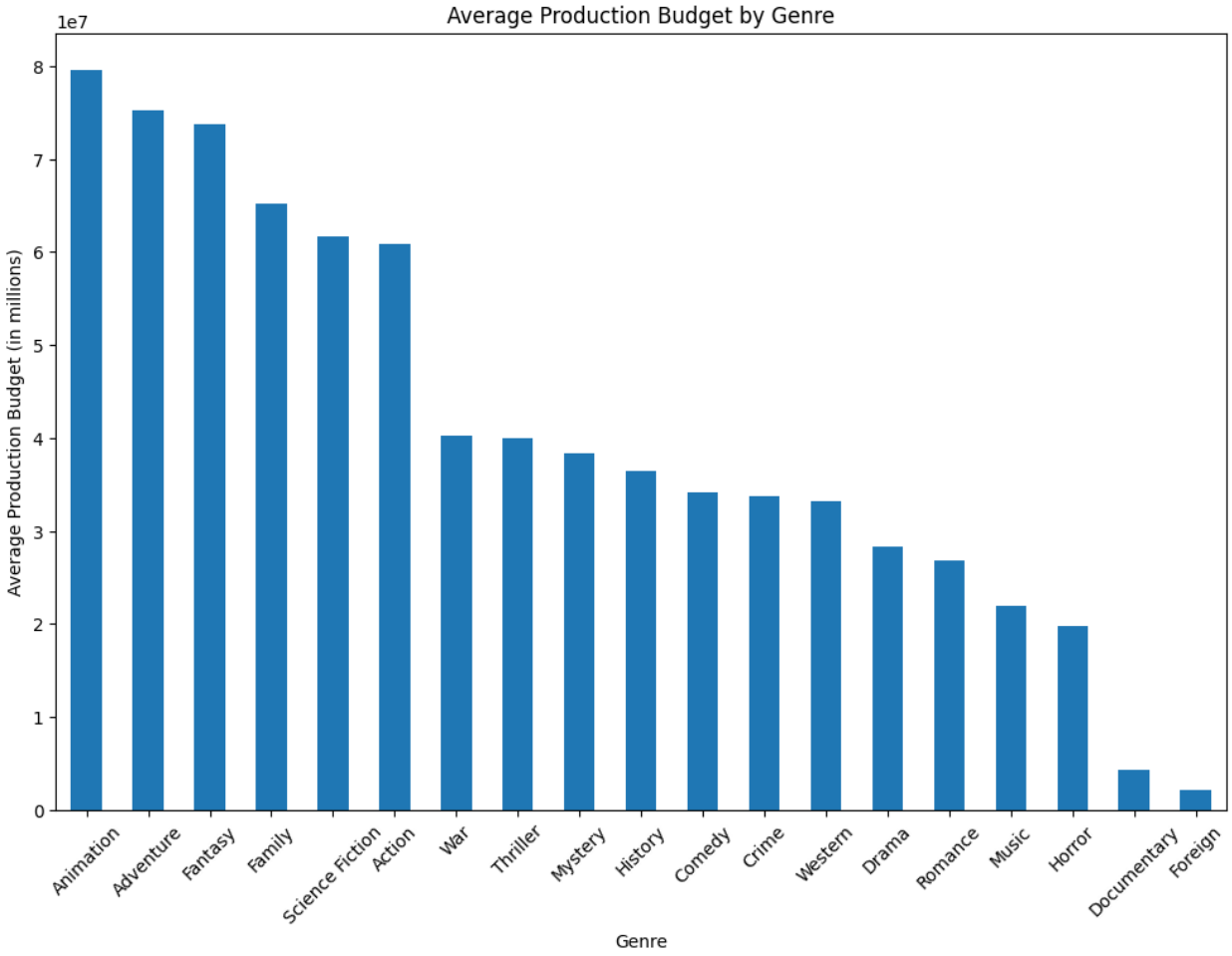
همانطور که مشاهده می‌شود rt_review_count با ۰/۷۶ و rt_production_budget با ۰/۷۱ قوی ترین پیش بینی کننده ها هستند. این به این معنی است که بودجه بیشتر و تعداد نظرات بیشتر مخاطبان باعث جذب و افزایش درآمد می‌شود. همچنین rt_audience_score (۰/۶۱) مهمتر از rt_critics_score (۰/۱۸) است که به این معنی است که آنچه مخاطبان عمومی فکر می‌کنند از نظرات منتقدان تأثیرگذارتر است.

برای آنالیز بهتر می‌توان میزان همبستگی هرکدام از این فیچرها را به صورت جداگانه plot کرد.

پاسخ سوالات EDA:

۱. **متوسط هزینه برای هر ژانر فیلم چقدر است؟**

برای رسیدن به جواب این سوال باید توزیع production budget را در ژانرها بررسی کنیم. برای این کار ابتدا نام ژانرها را از دیتاست استخراج می‌کنیم. سپس هر ژانر را با میانگین بودجه ای که برای تولید آن فیلم صرف شده است، groupby کرده و سورت می‌کنیم. خروجی به این صورت خواهد بود:



همانطور که مشاهده می‌شود بیشترین مقدار بودجه برای تولید انیمیشن و در جایگاه دوم برای ژانر ماجراجویی بوده است. همچنین می‌توانیم ۵ ژانری که بیشترین بودجه برای آنها صرف شده است را بدست آوریم.

```
top_5_average_expensive_genres = avg_budget_genre.head(5)
top_5_average_expensive_genres
```

✓ 0.0s Python

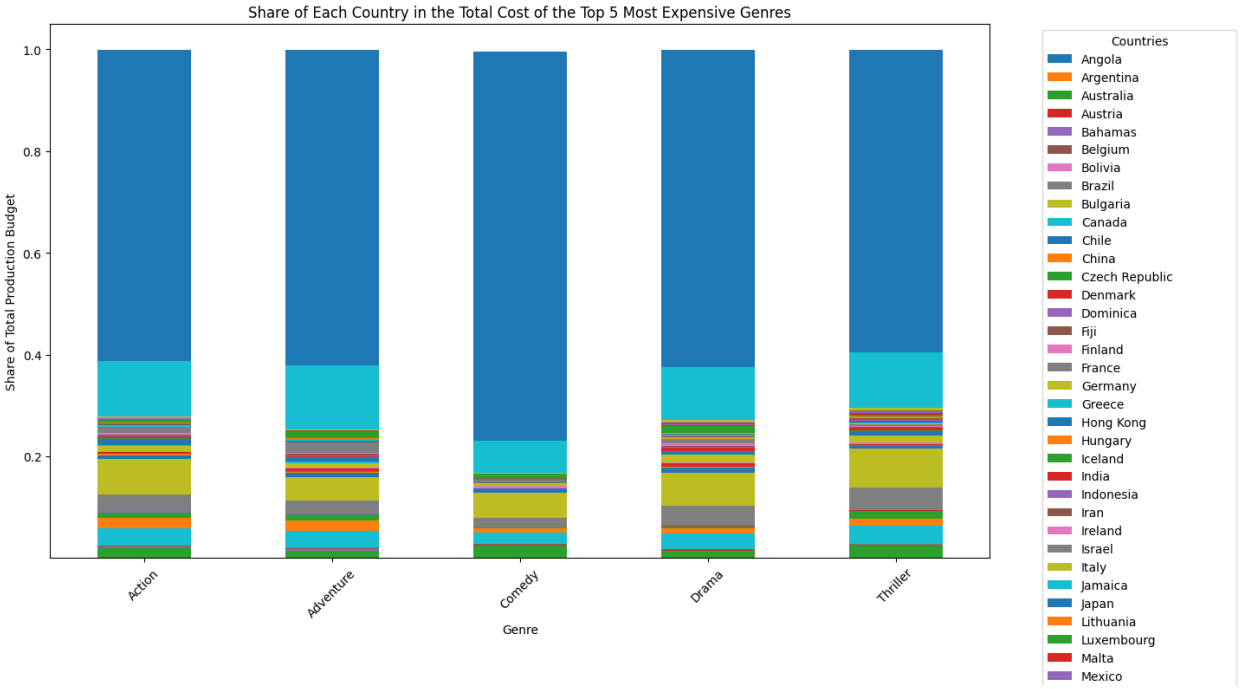
```
genres
Animation          7.955718e+07
Adventure          7.529758e+07
Fantasy            7.370141e+07
Family             6.520481e+07
Science Fiction    6.166937e+07
Name: rt_production_budget, dtype: float64
```

۲. **سهم هر کشور در مجموع هزینه هر ژانر فیلم چقدر است؟(برای ۵ تا از پر خرج ترین ژانر ها بدست بیاورید)**

برای رسیدن به پاسخ این سوال باید تجزیه و تحلیل کنیم که چگونه هزینه های تولید در بین کشورهای مختلف برای هر ژانر فیلم تقسیم می شود. به طور مشخص، می‌خواهیم سهم هر کشور از هزینه کل هر ژانر را محاسبه کنیم و پنج ژانر گران قیمت را شناسایی کنیم. بنابراین مراحل زیر باید طی شود:

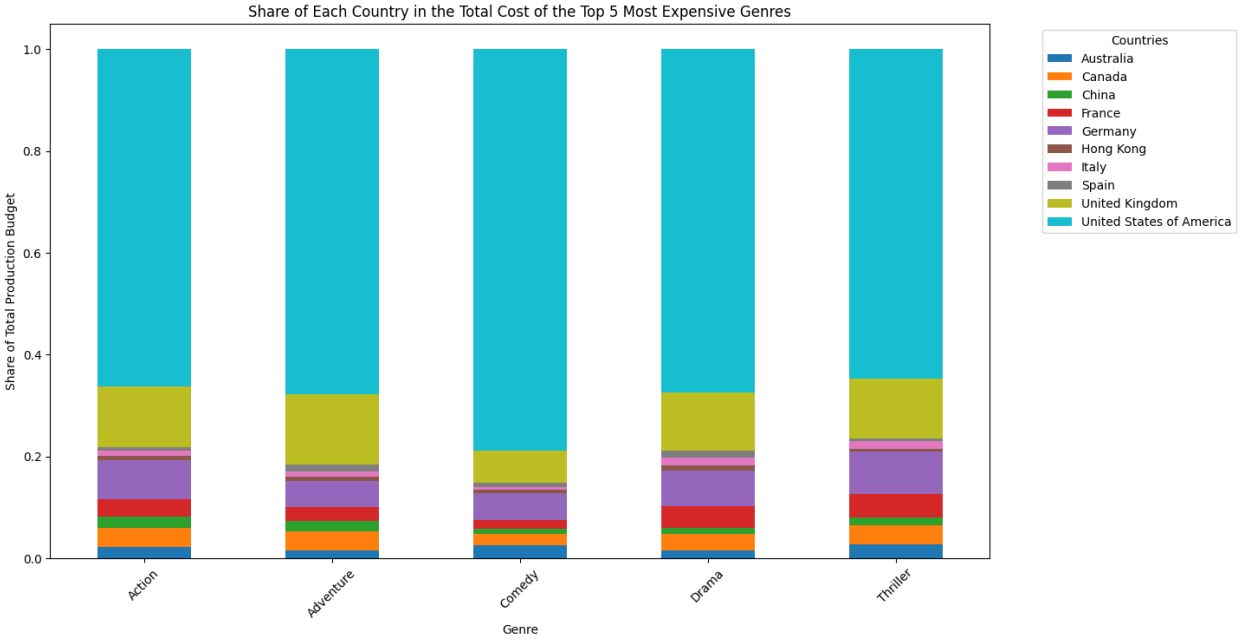
- استخراج نام کشورها: مشابه نحوه استخراج نام ژانرها، نام کشورها را از ستون rt_production_countries استخراج می کنیم.
- محاسبه کل هزینه تولید بر اساس ژانر: برای هر ژانر، کل هزینه تولید را محاسبه می کنیم.
- محاسبه سهم کشور بر اساس ژانر: برای هر کشور در هر ژانر، ما سهم آن را در کل هزینه تولید محاسبه خواهیم کرد.
- ۵ ژانر برتر را بر اساس هزینه کل تولید شناسایی خواهیم کرد.

خروجی به این صورت خواهد بود:



همانطور که می بینیم کشورهای زیادی وجود دارند و تجزیه و تحلیل داده ها امکان پذیر نیست. بنابراین من تجزیه و تحلیل را به ۱۰ کشور برتر که بیشترین فیلم را تولید می‌کنند محدود می‌کنم تا نتایج را بیشتر قابل مدیریت کند و روی مرتبط ترین داده ها تمرکز کند. بنابراین ابتدا باید بشماریم که هر کشور با چند فیلم مرتبط است. سپس ۱۰ کشور برتر را که بیشترین تعداد فیلم را دارند انتخاب می کنیم و در نهایت، میتوان تجزیه و تحلیل را برای آن ۱۰

کشور برتر ادامه داد.



همانطور که مشخص است ایالات متحده آمریکا بیشترین میزان سهم را دارد که با رنگ فیروزه ای مشخص شده است.
برای بررسی دقیق‌تر و داشتن مقادیر عددی این سهم‌ها، درصد هر یک را نیز محاسبه کردم.

genres	Action	Adventure	Comedy	Drama	Thriller
countries					
Australia	2.23	1.60	2.55	1.59	2.64
Canada	3.69	3.63	2.29	3.18	3.85
China	2.20	2.10	0.89	1.15	1.49
France	3.52	2.70	1.82	4.23	4.69
Germany	7.58	5.10	5.18	7.01	8.35
Hong Kong	0.82	0.90	0.76	1.08	0.54
Italy	1.12	0.97	0.46	1.58	1.51
Spain	0.71	1.33	0.79	1.39	0.49
United Kingdom	11.81	13.81	6.44	11.30	11.76
United States of America	66.33	67.85	78.81	67.49	64.67

مطابق با مشاهدات قبلی، کشور آمریکا بیشترین میزان سهم را در ۵ ژانر گران قیمت دارد و در جایگاه دوم نیز انگلیس قرار دارد.

نکته: همانطور که می بینیم ۵ ژانر برتر گران قیمت در اینجا متفاوت هستند! چیزی که اتفاق می افتد:

میانگین بودجه بر اساس ژانر: در پارت اول تحلیل خود، وقتی میانگین بودجه تولید بر اساس ژانر را محاسبه کردیم، وزن هر فیلم به یک اندازه بود و دیدم که ژانر انیمیشن بالاترین میانگین بودجه را دارد.

بودجه کل بر اساس ژانر: در پارت دوم، هنگام محاسبه کل بودجه تولید برای هر ژانر، بودجه تولید همه فیلم ها در هر ژانر را جمع می کنیم. ژانرهایی مانند اکشن و ماجراجویی به طور کلی می توانند فیلم های بسیار بیشتری داشته باشند، که با وجود پایین بودن میانگین بودجه برای آن ژانرها، بودجه کل تولید آنها را افزایش می دهد.

بنابراین ۵ ژانر گران در هر پارت متفاوت خواهد بود!

۳. **تعداد فیلم‌های ساخته‌شده را در ۱۰ سال گذشته مقایسه کنید.**
برای پاسخ به این سوال، باید فیلم‌های ۱۰ سال گذشته را فیلتر کنیم، سپس تعداد فیلم‌های ساخته شده در ژانر گران قیمت را در آن دوره با هم مقایسه کنیم. برای این کار ابتدا فیلم‌های ۱۰ سال گذشته را بر اساس ستون rt_release_date فیلتر میکنیم. با استفاده از سوال قبلی که در آن میانگین هزینه تولید به تفکیک ژانر را پیدا کردیم، ژانرهای گران قیمت را شناسایی میکنیم. و در نهایت تعداد فیلم های ساخته شده در آن ژانرها را در ۱۰ سال گذشته را میشماریم.
خروجی به این صورت خواهد بود:

همانطور که مشاهده می‌شود هر ژانر به همراه تعداد آن در این لیست ثبت شده است.

```
Adventure          675
Action             462
Fantasy            351
Comedy             307
Family             292
Science Fiction    230
Thriller           221
Drama              207
Animation          191
Romance             88
Crime               53
Horror              51
Mystery             45
War                 30
History             23
Western             19
Music               17
Documentary          1
Foreign              1
Name: genres, dtype: int64
```

۴. **به طور متوسط کدام کشورها طولانی‌ترین و کوتاه‌ترین فیلم‌ها را می‌سازند؟**
برای پاسخ به این سوال، باید میانگین زمان پخش فیلم ها را برای هر کشور محاسبه کنیم. ابتدا ستون کشورها را explode می کنیم تا هر کشور ردیف خودش را بگیرد. سپس بر اساس کشور و میانگین زمان اجرای هر کشور

گروهبهندی می‌کنیم. در نهایت برای تعیین طولانی ترین و کوتاه ترین فیلم ها، کشورها را بر اساس میانگین زمان پخش مرتب میکنیم. برای این کار من ۱۰ تا از کشورها را ملاک قرار داده ام.

لیست بالا شامل کشورهایی با بیشترین میانگین زمان فیلم و لیست زیر شامل کشورهایی با کمترین میانگین زمان فیلم است.

```
Top 10 countries with the longest average runtime:
production_countries_list
Singapore      172.00
Dominica       151.00
Slovenia       150.00
Malta          143.00
Morocco        141.33
Portugal       133.00
Jamaica        130.50
New Zealand    130.45
Philippines    130.00
Italy          128.67
Name: rt_runtime, dtype: float64

Top 10 countries with the shortest average runtime:
production_countries_list
Monaco      84.00
Angola      85.00
Iran        89.00
Israel      90.00
Fiji        98.00
Bolivia     98.00
Bulgaria    98.33
Thailand    98.67
Greece      99.00
Peru        99.50
Name: rt_runtime, dtype: float64
```

۵. **به غیر از انگلیسی، پرتکرارترین زبانها در فیلمها چه هستند؟**

برای پاسخ به این سوال ابتدا باید نام زبان ها را استخراج کنیم. سپس انگلیسی را از لیست زبانها حذف کرده و تعداد دفعات تکرار هر زبان را بشماریم و آنها را مرتب کنیم

همانطور که می‌بینیم پرکاربردترین زبان ها به غیر از انگلیسی فرانسه و رتبه دوم اسپانیایی است.

```
Most frequently used languages (apart from English):

Français     316
Español      281
Deutsch      184
Русский      153
Italiano     146
日本語           83
普通话           82
56       العربية
Português     49
Latin         43
Name: languages_list, dtype: int64
```

۶. **آمریکا در ۱۰ سال گذشته، به طور متوسط در هر سال چقدر در صنعت فیلمسازی هزینه کرده است؟ (به تفکیک سال)**

برای پاسخ به این سوال ابتدا داده ها را فیلتر میکنیم تا فقط فیلم‌های تولید شده در "ایالات متحده آمریکا" را شامل شود. سپس داده های ۱۰ سال گذشته را بر اساس ستون rt_release_date فیلتر میکنیم. بعد داده ها را بر اساس سال گروه بندی کرده و بودجه تولید را برای هر سال جمع میکنیم. در آخر میانگین بودجه تولید ۱۰ سال گذشته را محاسبه میکنیم.

خروجی:

```
United States total spending on the film industry:
release_year
2015    6,107,500,003$
2016    4,475,400,000$
Name: rt_production_budget, dtype: object
```

از آنجایی که آخرین رکورد تولید فیلم ۲۰۱۶ است، فقط داده های دو سال قابل نمایش است. برای اصلاح این موضوع سال جاری را به ۲۰۱۶ تغییر می‌دهم تا داده های ۲۰۰۶ تا ۲۰۱۶ قابل مشاهده باشد.

```
United States total spending on the film industry:
release_year
2006    6,065,050,000$
2007    5,250,715,000$
2008    6,341,900,000$
2009    6,758,546,652$
2010    7,052,202,650$
2011    6,926,255,000$
2012    6,871,525,000$
2013    7,522,270,010$
2014    6,855,000,000$
2015    6,107,500,003$
2016    4,475,400,000$
Name: rt_production_budget, dtype: object
```

۷. **روند قبلی را بدون در نظر گرفتن کشور برای ۱۰ سال گذشته مقایسه کنید.**

برای این سوال هم همین کار را بدون در نظر گرفتن و فیلتر کردن کشور خاصی انجام میدهیم که خروجی در زیر قابل مشاهده است. با پاسخ این سوال متوجه خواهیم شد که تقریبا ۹۰ درصد از هزینه های مربوط به ۱۰ سال گذشته فقط برای آمریکا بوده است.

```
Total spending on the film industry each year (all countries):
release_year
2006    6,466,729,867$
2007    5,637,827,510$
2008    6,836,899,099$
2009    7,132,277,844$
2010    7,428,370,336$
2011    7,552,445,771$
2012    7,108,572,154$
2013    7,914,389,300$
2014    7,017,500,000$
2015    6,394,500,003$
2016    4,593,390,000$
Name: rt_production_budget, dtype: object
```

۸. Johnny Depp در چه فیلمهایی بازی کرده است؟

برای اینکه بفهمیم جانی دپ در کدام فیلم‌ها بازی کرده است، باید df_credit DataFrame را بر اساس نام او در ستون rt_actors فیلتر کنیم.

خروجی:

```
Movies starring Johnny Depp:
 117     Charlie and the Chocolate Factory
178                                 Rango
333                          Transcendence
499                          Jack and Jill
783                              Mortdecai
1119                        21 Jump Street
1203                         Secret Window
1378           A Nightmare on Elm Street
1581                                 Blow
1594                          Corpse Bride
1701           Once Upon a Time in Mexico
1890                     Don Juan DeMarco
2051                        The Libertine
2108                  Edward Scissorhands
2310       Fear and Loathing in Las Vegas
3939                                 Tusk
Name: rt_title, dtype: object
```

۱۰. به طور متوسط چند درصد نقش اول تا پنجم فیلم ها(به تفکیک برای هر نقش) مرد، و چند درصد زن هستند؟

برای محاسبه درصد بازیگران زن و مرد در نقش های اول تا پنجم فیلم ها باید:۵ بازیگر برتر هر فیلم را از ستون rt_actors در df_credit استخراج کنیم. سپس جنسیت آنها را بررسی کرده (با فرض اینکه ۱ = زن، ۲ = مرد) و درصد مردان و زنان را به طور جداگانه برای هر نقش محاسبه کنیم (اول تا پنجم).

```
Percentage of Male and Female Actors in the First to Fifth Roles:
             Male     Female
Role 1  66.720322  25.674044
Role 2  52.888889  36.606061
Role 3  52.270884  34.590430
Role 4  52.625103  32.731747
Role 5  54.510961  28.583474
```

همانطور که مشاهده می‌شود اکثر بازیگران نقش اول مرد هستند.

۱۱. محبوب ترین ژانرهای فیلم در ۱۰ سال گذشته به چه ترتیب بوده است؟(یکبار بر اساس تعداد review و یکبار بر اساس critics_score مقایسه کنید)

ابتدا فیلم های ۱۰ سال گذشته را فیلتر میکنیم. سپس ژانرها را explode می کنیم تا هر ژانر در یک ردیف جداگانه باشد. مرحله بعد گروه بندی بر اساس ژانرها و مجموع تعداد کل نظرات و گروه بندی بر اساس ژانرها و میانگین امتیاز منتقدان است.

```
Most Popular Genres (Based on Number of Reviews):
rt_genres
Action             838917
Adventure          725011
Drama              687553
Thriller           578257
Comedy             505399
Science Fiction    502029
Fantasy            336038
Crime              264444
Family             261666
Romance            205042
Name: rt_review_count, dtype: int64
```

```
Highest Rated Genres (Based on Average Critics Score):
rt_genres
Documentary    6.80
War            6.76
History        6.67
Foreign        6.57
Drama          6.52
Western        6.51
Romance        6.36
Animation      6.34
Mystery        6.32
Music          6.27
Name: rt_critics_score, dtype: float64
```

همانطور که می بینیم محبوب ترین ژانر بر اساس تعداد نظر کابران اکشن و بر اساس امتیاز منتقدان، مستند است.

پیش‌بینی درآمد فیلم‌ها

آماده‌سازی داده‌ها

پس از اطمینان از تمیز بود دیتا و مراحلی که در پیش پردازش انجام داده ایم، مرحله بعدی می‌تواند ایجاد فیچرهای جدید باشد.

ایجاد ویژگی‌های جدید

من این فیچرها را اضافه کرده ام:

۱. **تعداد بازیگران مشهور**: فیلم هایی که بازیگران شناخته شده‌ای دارند ممکن است مخاطبان بیشتری را جذب کنند که می‌تواند باکس آفیس را افزایش دهد. برای این کار لیستی از بازیگران مشهور را ایجاد کرده و تعداد بازیگران را به ازای هر فیلم میشماریم. این فیچر با نام num_famous_actors به دیتا اضافه می‌شود.

```
# List of famous actors
famous_actors = ['Johnny Depp', 'Leonardo DiCaprio', 'Meryl Streep', 'Brad Pitt',
'Scarlett Johansson', 'Will Smith', 'Tom Hanks', 'Angelina Jolie']

# Function to count number of famous actors in the cast
def count_famous_actors(actors_list):
    if isinstance(actors_list, list):
        return sum(1 for actor in actors_list if actor['name'] in famous_actors)
    return 0

df_credit['num_famous_actors'] =
df_credit['rt_actors'].apply(count_famous_actors)
```

۲. **محبوبیت کارگردان**: کارگردانانی که به خاطر فیلم‌های موفق خود شناخته می‌شوند یا شهرت بالایی دارند، می توانند شانس موفقیت یک فیلم را افزایش دهند. برای این کار نیز لیستی از کارگردانان معروف را ایجاد کرده و یک ویژگی باینری ایجاد میکنیم که نشان دهد کارگردان فیلم در لیست است یا خیر.

```
# List of famous directors
famous_directors = ['Steven Spielberg', 'Christopher Nolan', 'Quentin Tarantino',
'Martin Scorsese', 'James Cameron']

def is_famous_director(directors_list):
    if isinstance(directors_list, list):
        for director in directors_list:
            if director['name'] in famous_directors:
                return 1
    return 0
df_credit['is_famous_director'] = df_credit['rt_staff'].apply(is_famous_director)
```

۳. **تعداد ژانرها**: ترکیب چندین ژانر ممکن است برای مخاطبان وسیع تری جذاب باشد. برای این فیچر نیز کافیست تعداد ژانرهای مرتبط با هر فیلم را بشماریم.

```
def count_genres(genres_list):
    if isinstance(genres_list, list):
        return len(genres_list)
    return 0

df_movies['num_genres'] = df_movies['rt_genres'].apply(count_genres)
```

آموزش مدل‌ها

۱. ابتدا از **رگسیون خطی** برای آموزش مدل استفاده می‌کنیم. (در ابتدای کد تارگت ما دراپ شده است بنابراین نگرانی از این بابت وجود ندارد.)

```
Linear Regression - MSE: 1.6798627295488442e+16
Linear Regression - R2: 0.6745484509116488
```

همانطور که قابل مشاهده است، مقدار MSE = 1.67 و R2 score = 0.67 است.

۲. **Random Forest Regressor:** خروجی این مدل نیز به این صورت است:

```
Random Forest Regressor - MSE: 1.4251111750597056e+16
Random Forest Regressor - R2: 0.723903250314468
```

۳. **Gradient Boost Regressor:**

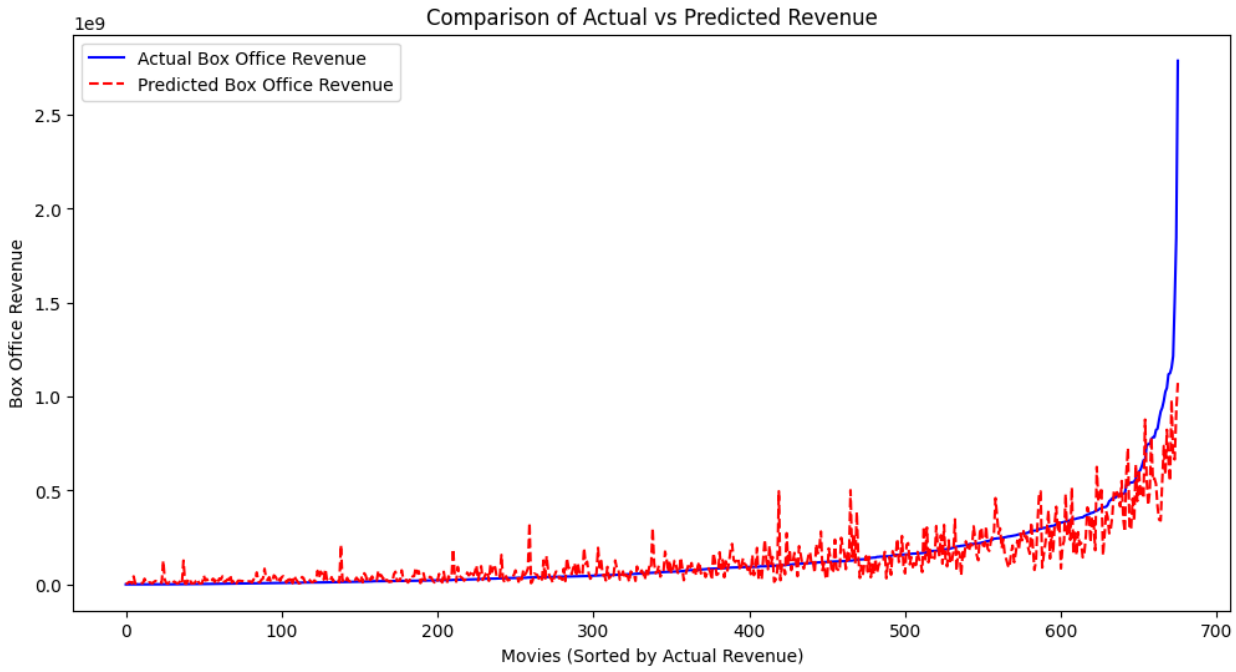
```
Gradient Boosting Regressor - MSE: 1.3714187216526852e+16
Gradient Boosting Regressor - R2: 0.734305464631326
```

Hyperparameter Tuning for Random Forest:

تنظیم Hyperparameter به یافتن تنظیمات بهینه برای مدل Random Forest برای بهبود عملکرد کمک می کند. به جای استفاده از مقادیر پیش‌فرض، پارامترهایی مانند تعداد درختان، عمق و تقسیم‌ها را بهینه می‌کنیم. ما از RandomizedSearchCV برای جستجوی بهترین ترکیب ابرپارامترها استفاده می کنیم و پس از یافتن بهترین هایپرپارامترها، مدل بهینه شده را آزمایش می کنیم.

چارت‌ها:

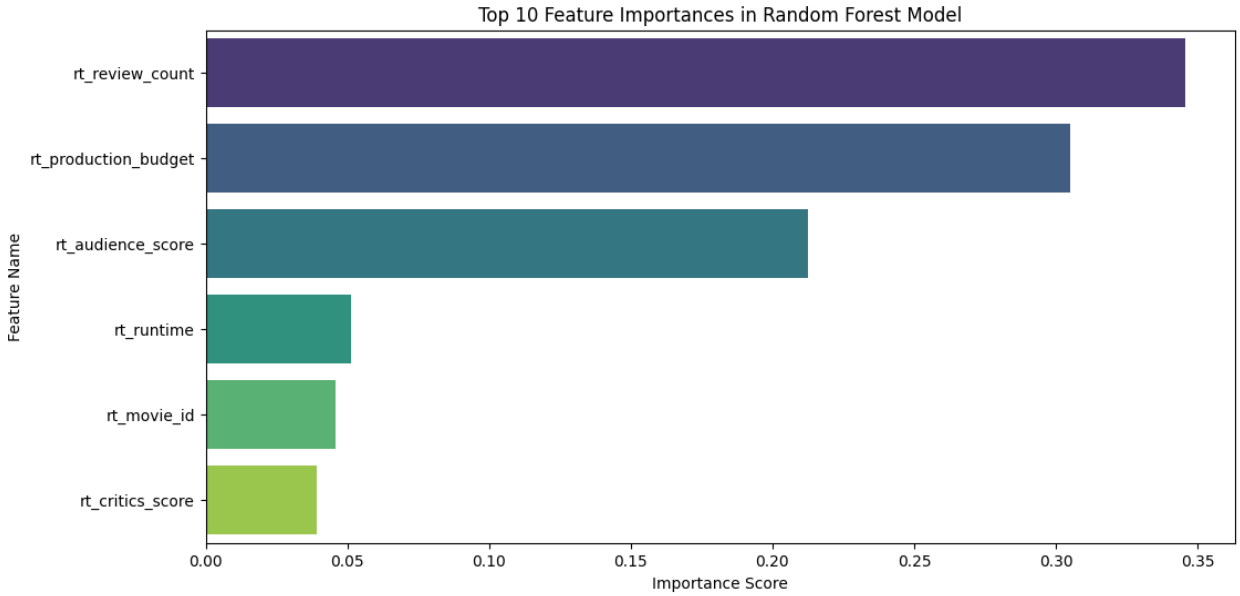
برای نشان دادن اینکه چقدر پیش‌بینی‌های مدل با مقادیر واقعی مطابقت دارند، می‌توانیم از نمودار خطی استفاده کنیم که در آن: محور x نمایانگر شاخص نمونه (فیلم) و محور y درآمد باکس آفیس را نشان می دهد. ما هم مقادیر واقعی و هم مقادیر پیش بینی شده را رسم می کنیم تا روند آنها را مقایسه کنیم.



همانطور که مشاهده می‌شود مقدیر پیش‌بینی تا حد خوبی نزدیک به واقعیت هستند و این نشان‌دهنده عملکرد خوب مدلها است.

:Feature Importance

اهمیت ویژگی به درک اینکه کدام عوامل بیشتر در پیش بینی درآمد کمک می کنند کمک‌کننده است. ما می‌توانیم اهمیت ویژگی‌ها را از مدل رندوم فارست ترین شده استخراج کنیم و با استفاده از نمودار میله‌ای آن‌ها را visualize کنیم.



همانطور که مشاهده می‌شود، review_count، production_budget و audience_score بیشترین میزان تاثیر در پیش‌بینی را داشته اند. بنابراین اینها عوامل کلیدی تاثیرگذار بر درآمد فیلم ها هستند.

نتیجه‌گیری

از میان سه مدل آموزش‌دیده شده می‌توان گفت رندوم فارست عملکرد بهتری خواهد داشت. این مدل از رگرسیون خطی برای پیش‌بینی درآمد فیلم بهتر عمل می‌کند، زیرا به طور موثر ویژگی‌های غیرخطی، دسته‌بندی و عددی و داده‌های از دست رفته را بدون نیاز به پیش‌پردازش گسترده مدیریت می‌کند. برخلاف رگرسیون خطی، که یک رابطه خطی بین متغیرها را فرض می‌کند، این روش الگوهای پیچیده را با ساختن درخت‌های تصمیم‌گیری متعدد و میانگین‌گیری خروجی‌های آنها، باعث کاهش اورفیت و بهبود تعمیم میشود. همچنین همانطور که مشاهده شد feature importance را ارائه می دهد که به شناسایی تأثیرگذارترین عوامل مؤثر بر درآمد کمک می کند.

همچنین برای بهبود بیشتر عملکرد پیش‌بینی، می‌توانیم ویژگی‌های مرتبط‌تر را مهندسی کنیم، مانند تجزیه و تحلیل احساسات از توضیحات فیلم، نظرات رسانه‌های اجتماعی، یا میزان علاقه مخاطبان قبل از انتشار فیلم. همچنین تکنیک‌های انتخاب ویژگی و کاهش ابعاد مانند PCA هم می‌تواند به حذف داده‌های outlier و بهبود کارایی مدل کمک کند.