CİHAT BERA ŞİMŞEK

30551

DSA 210 – Introduction to Data Science Term Project

Addressing My Own Movie Taste

Introduction

Letterboxd is a platform getting popular among young people especially for the ability to be used as a movies social media. I have decided to use my letterboxd ratings to analyze my personal movie taste. There is a total of 245 movies rated in my account. Dataset contains *movie name, release date, budget, box office revenues, production company, duration, director, country of director, genre, subgenre, personal rating, imdb rating* and *rateyourmusic.com rating* of that movie. RYM is chosen for a specific reason because it's a niche website with lower counts of rating. IPYN and data.json file is on my github.

Visualization Techniques

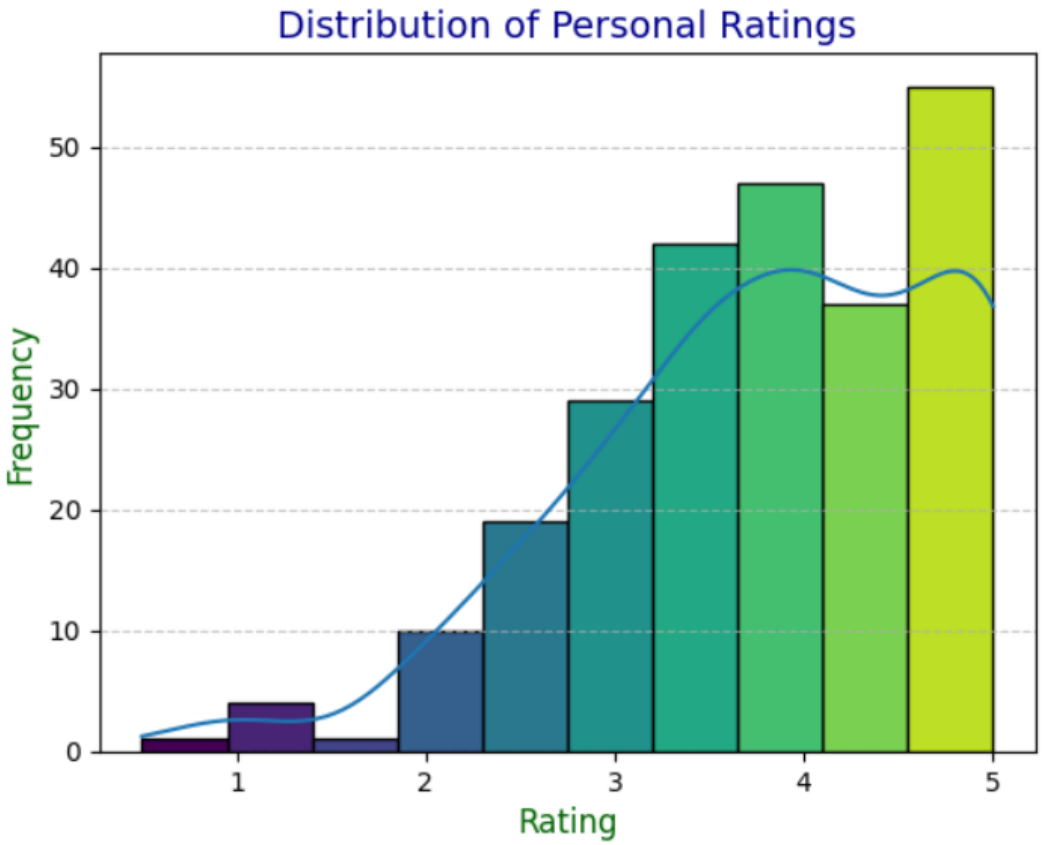
In this project, I have used *bar charts* to **visualize numerical data** trends, such as average box office revenues by genre and the number of movies released by year. Also I have used *scatter plots* for revealing relationships between features like budget and box office, *pie charts* for **representing categorical distributions,** e.g., director countries, while feature importance for identifying influential factors. Most of the data had to be transformed with logarithmic scale due to dense and close numbers.

Future of the Project

After taking DSA 210 course I want to expand my movies database much more and actually be able to guess the next movies I'm going to watch by feeding the current movie theater data into my own models. I also want to be able to guess IMDB and RYM ratings of particular movies.

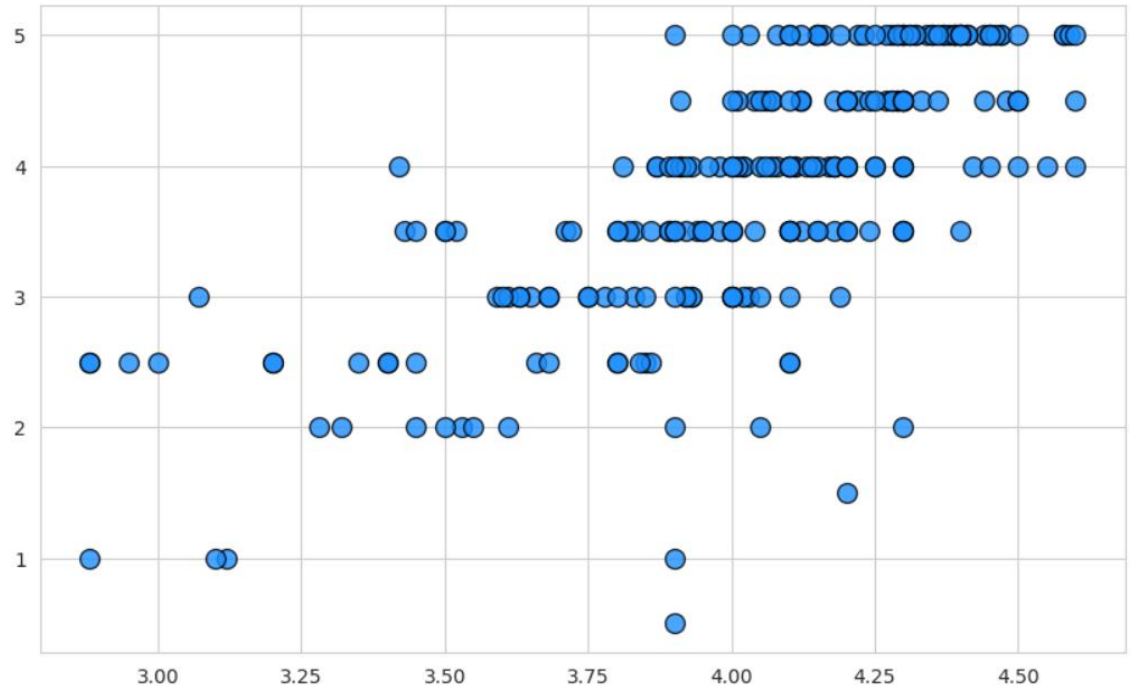
Data Visualization & Analyzing Some Hypothesis

On their own, my ratings are generally positive and skewed toward ratings of 3 or higher, indicating a selective movie-choosing behavior. Normally, we would expect a normal distribution where extreme ratings of 1 and 5 are less frequent, with most ratings clustering around the middle. However, the lack of balance in the distribution and its deviation from a normal curve suggest a systematic bias influencing the ratings. This deviation challenges the hypothesis that my ratings are entirely objective and unaffected by external factors. The figure below illustrates the distribution of my personal ratings.

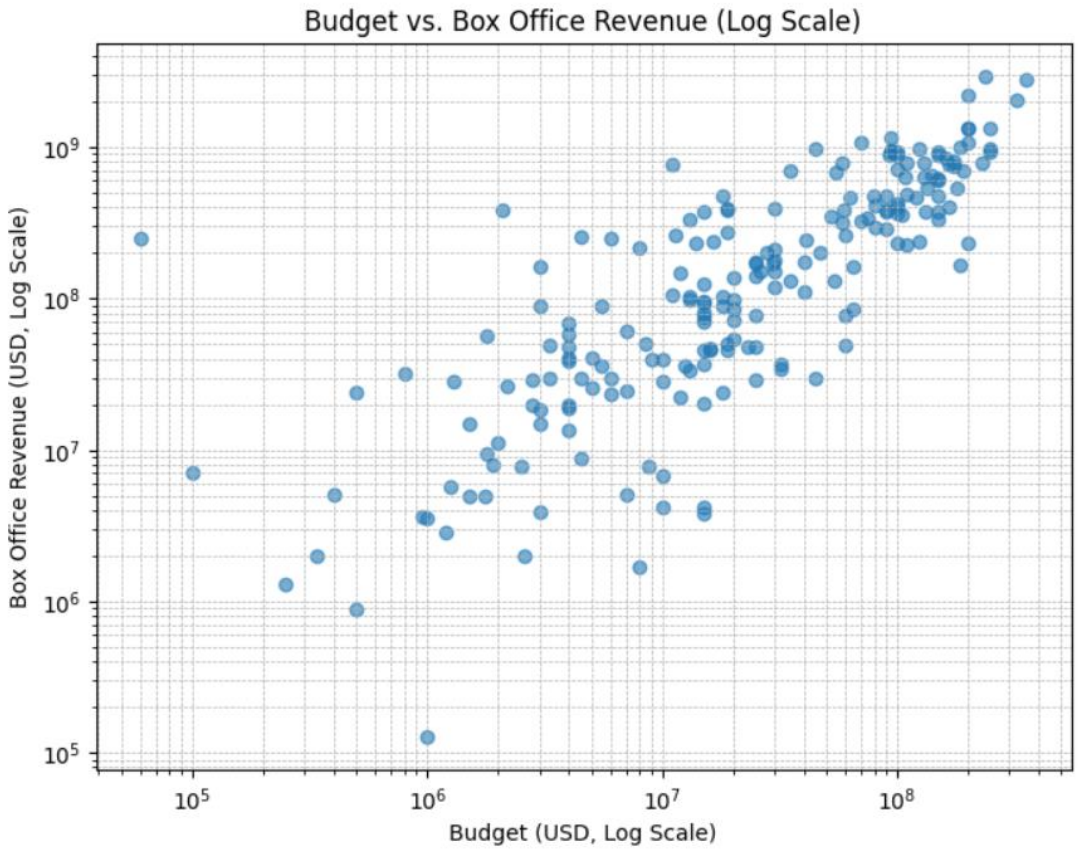


Above is the figure of RYM ratings vs personal ratings. A dense overlap can be seen between 3.75 - 4.25 segment. This can mean RYM is a source of affect on my ratings. This is a hypothesis worth to test.

The graph below personal rating vs rym ratings also seems to support this is worth testing because there is a density between 5 scores on x label personal ratings and over 4 label on y label, rym ratings.

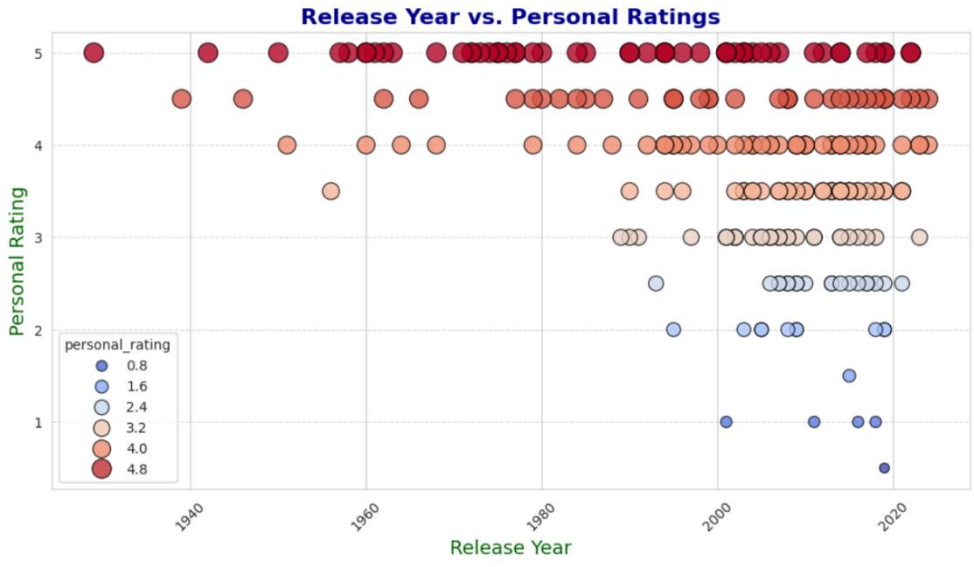


But before testing, let's analyze the data from different point of views.

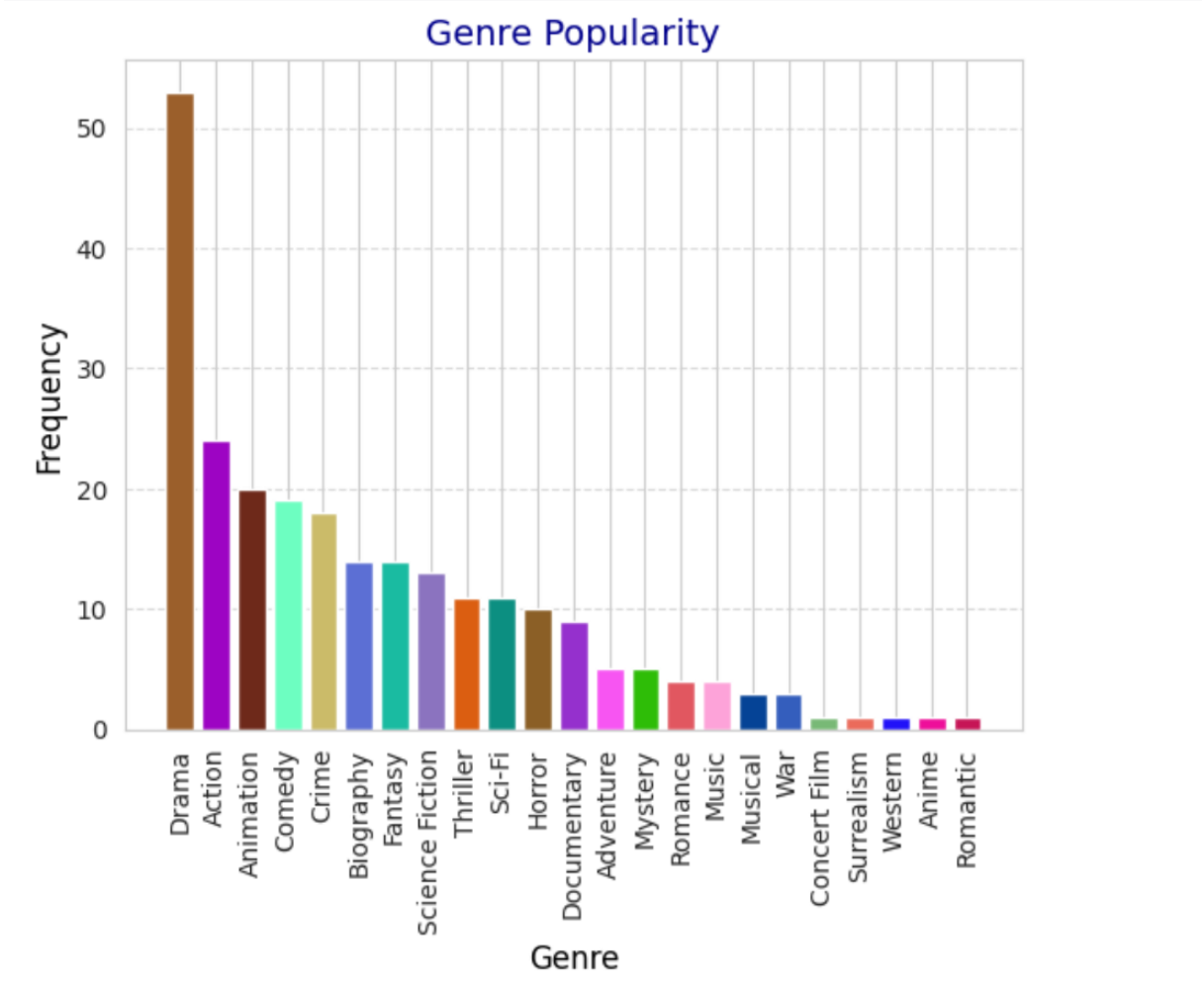


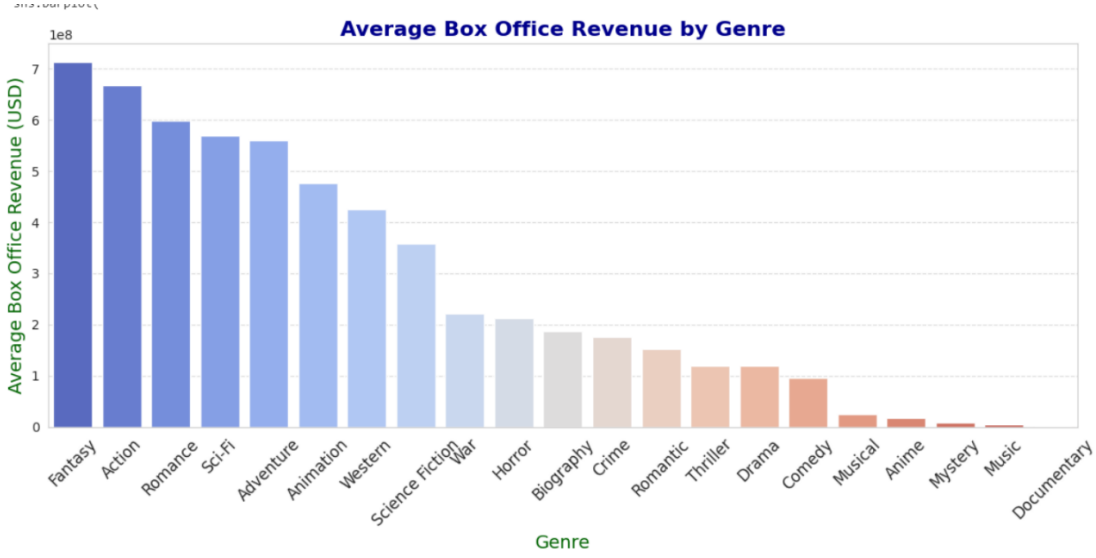
The data points generally follow an upward trend, suggesting that movies with larger budgets tend to generate higher box office revenue. Data aligns with the industry norm where bigger budgets allow for higher earnings. It can be said that there's not much movie where producers put lower than 10^7 USD on budget on logscale and gained much from it. Gaining revenues lower or higher than the budget seems equally probable when budget is lower than 10^7. It seems that spending over 10^7 USD on budget gets at least the budget revenue most of the time and most probably more than the budget.

I seem to favor movies close to years 1960, 1980 and 2000. Leaving aside the probable implication on chart that really good movies appear every 20 years, release year can be significant on personal ratings. (Note that this is tested with chi square and saw a minimal p-value, shown just before random forest test after duration chi-square.)

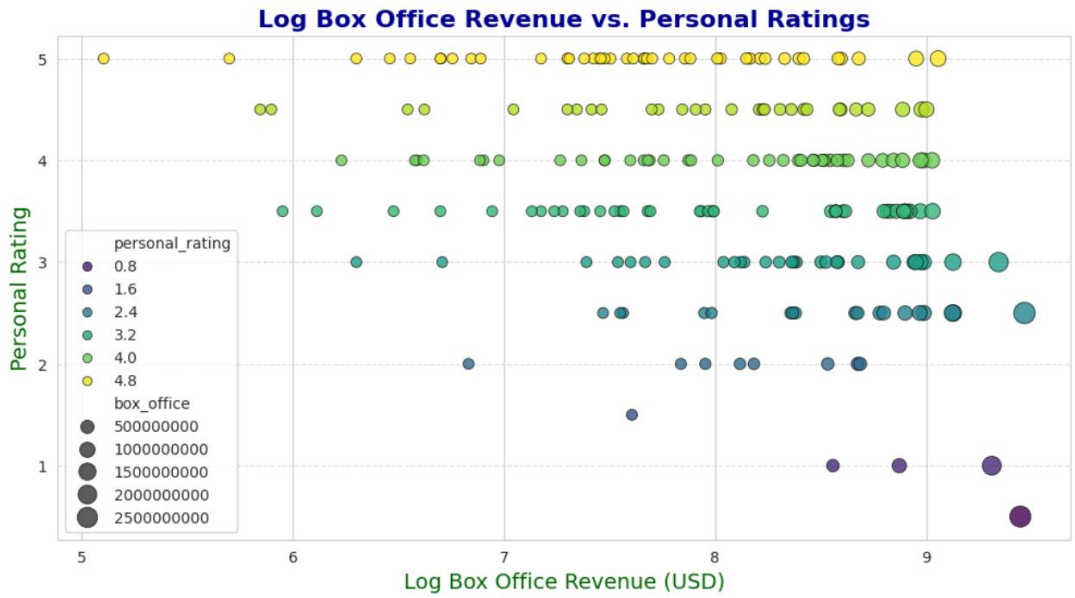


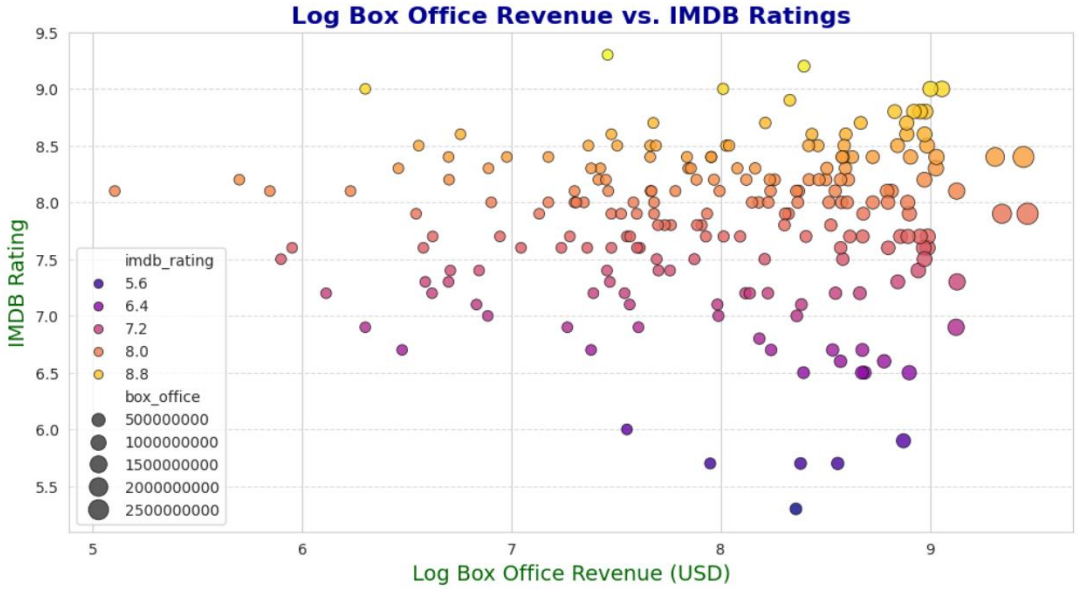
Most chosen movie genre by me is drama, followed by action and animation.

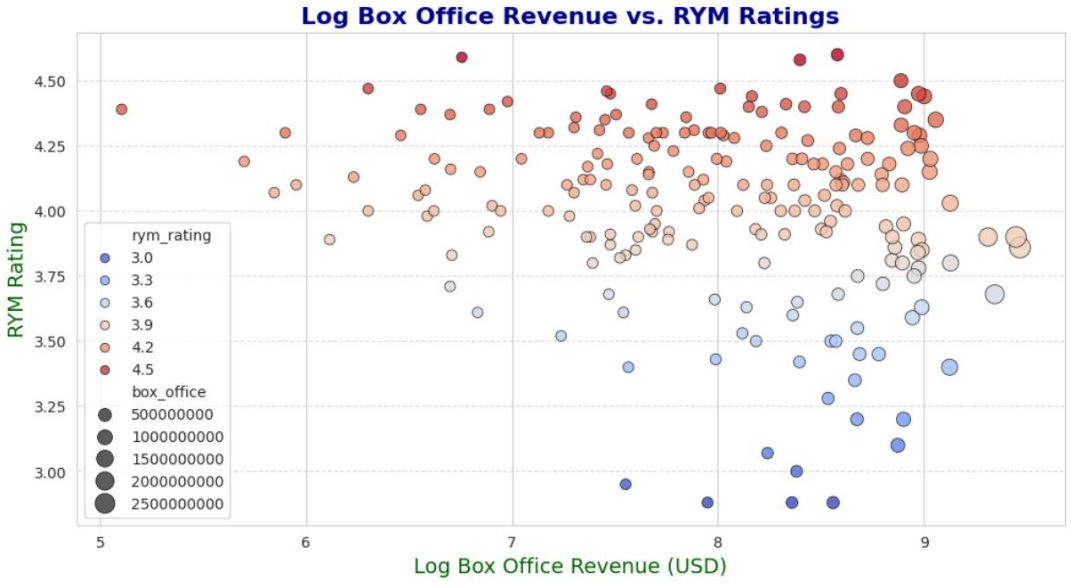




The most revenue bringing genre is fantasy followed by action and romance. There are some overlaps on top 10 in both charts. Can revenue be an affective parameter on my personal choices and public ratings?


Log Box Office Revenue vs. Personal Ratings
Personal Rating
Log Box Office Revenue (USD)
personal_rating
0.8
1.6
2.4
3.2
4.0
4.8
box_office
500000000
1000000000
1500000000
2000000000
2500000000


Log Box Office Revenue vs. IMDB Ratings
IMDB Rating
Log Box Office Revenue (USD)
imdb_rating
5.6
6.4
7.2
8.0
8.8
box_office
500000000
1000000000
1500000000
2000000000
2500000000


Log Box Office Revenue vs. RYM Ratings
RYM Rating
Log Box Office Revenue (USD)
rym_rating
3.0
3.3
3.6
3.9
4.2
4.5
box_office
500000000
1000000000
1500000000
2000000000
2500000000

```
Chi-square test for Revenue vs. Personal Ratings: Chi2=19.09593789206504, p-value=0.0007525556417063393
Chi-square test for Revenue vs. IMDB Ratings: Chi2=7.341625959117045, p-value=0.1188991246394403
Chi-square test for Revenue vs. RYM Ratings: Chi2=7.777718692867435, p-value=0.10006849948510817
```

Above is the chi square tests for revenues affect on ratings. Revenues seems statistically significant on personal ratings but not on public ratings.

```
Chi-square test for IMDB Ratings vs. Personal Ratings: Chi2=81.05875656785652, p-value=1.0391304848796049e-16
```

```
Chi-Square Test Results:
Chi-Square Statistic: 78.2554
P-Value: 0.0000
Degrees of Freedom: 2
Expected Frequencies:
[[ 6.32098765 53.72839506 35.95061728]
 [ 9.67901235 82.27160494 55.04938272]]

Result: The relationship between RYM Rating and Personal Rating is statistically significant.
```

Although there is no revenue significancy on public ratings both public ratings have much much lower p values (so much that on second test it's rounded to 0). But relationship that causes this significancy between public and personal ratings is not revenues. So we are not going to include revenues on random forest. Both public ratings are so low on p values we have to compare their significancy with other methods. Both being so significant on personal rating, which one is the most important feature? We have to run a random forest to learn. But first let's check other features chi-square and leave the feature importance at the end.

```
Chi-square test for Duration vs. Personal Ratings:
Chi2 = 3.857335893879396
p-value = 0.4256574973586905
```

Duration seems not significant on personal ratings. This is another hypothesis refuted. We don't have to include duration to our random forest.

```
Chi-square test for Budget vs. Personal Ratings:
Chi2 = 29.650054568184277
p-value = 5.766561444289418e-06
```

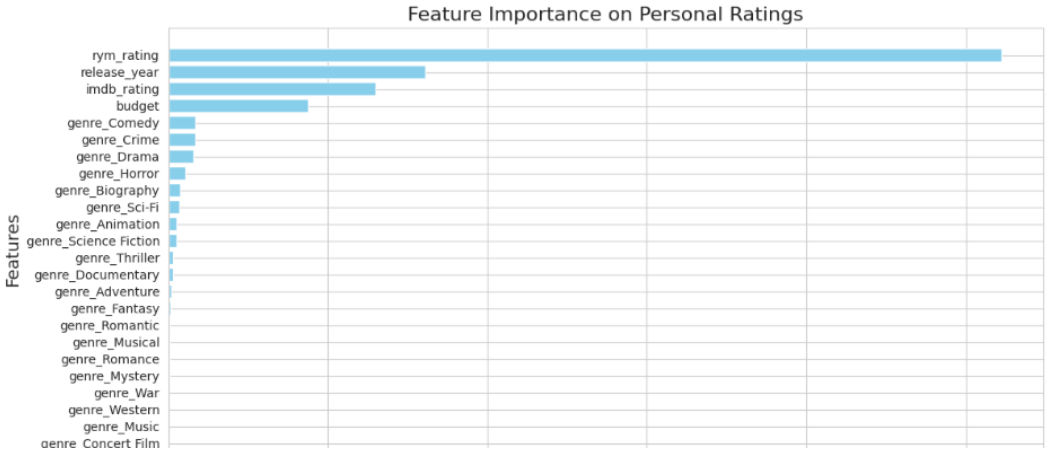
```
Chi-square test for Release Year vs. Personal Ratings:
Chi2 = 47.29976406726752
p-value = 1.3206921010073579e-09

Chi-square test for Genre vs. Personal Ratings:
Chi2 = 64.22153060873914
p-value = 0.008895212264011259
```

Genre, release year and budget is statistically significant on personal rating. Now let's see the importance of features. We will race the significant ones with random forest.

```
Mean Squared Error on Test Set: 0.5542073863636363

Feature Importances:
                Feature  Importance
1            rym_rating    0.522587
3          release_year    0.161346
2           imdb_rating    0.129750
0                budget    0.087436
7          genre_Comedy    0.017143
9           genre_Crime    0.016851
11          genre_Drama    0.016002
13          genre_Horror    0.010747
6       genre_Biography    0.007409
19        genre_Sci-Fi    0.006901
5       genre_Animation    0.005298
20  genre_Science Fiction    0.005043
21       genre_Thriller    0.003259
10     genre_Documentary    0.003244
4       genre_Adventure    0.001870
12        genre_Fantasy    0.001583
18       genre_Romantic    0.001092
15        genre_Musical    0.000736
17        genre_Romance    0.000675
16        genre_Mystery    0.000546
22            genre_War    0.000426
23        genre_Western    0.000043
14          genre_Music    0.000015
8     genre_Concert Film    0.000000
```



Conclusion

Finally, based on related random forest result above, it is evident that the RYM (Rate Your Music) rating is the most influential factor in determining my personal ratings, significantly outweighing other features. The hypothesis that my ratings are entirely objective is refuted implicitly,

as they are clearly shaped by external factors, particularly the RYM rating. We can also say that IMDB rating is not important as it seems because there is a high significance relationship between budget and revenues and there is no significance between revenues and IMDB rating. This implies IMDB ratings may not be as impactful in shaping personal preferences as the random forest result suggests. Visualized feature also refutes the minimal effects of genre-specific factors rounding most of their contribution to almost zero.