Analysis of Movie Box Office Trend by Genre

Name: Isabella Yoo
USC ID: 13005966908
GitHub: IsabellaYoo

Introduction

Many industries are impacted by economic situations around the world, and the film industry is no exception. According to research conducted in South Korea, there is a significant correlation between the economic cycle and box office performance of specific movie genres. The researcher found that fantasy and action genres are directly proportional to the economic trends, while comedy and romance genres are inversely correlated (Lee, 2019). Building on this insight, this project aims to determine whether these patterns apply to the U.S. film market. Furthermore, this project will analyze volatility and interrelationship between the box office performance of different genres to study general patterns in genre popularity.

Data Collection

The data used in this study is from two sources: Federal Reserve Bank of St. Louis (FRED) and The Numbers. The Real GDP growth rate, used as the economic cycle indicator, comes from the FRED API. The data covers an annual period from 1930 to 2023, comprising a total of 94 data. Movie genre popularity, as determined by box office success, is obtained from market share percentage data per genre, which is gathered through web scraping from The Numbers. The Numbers distinguished movie genres into seven categories: Adventure, Action, Comedy, Drama, Thriller/Suspense, Horror, Romantic Comedy. For each genre, The Numbers calculated annual market share percentage from 1995 to 2024. Thus, 210 data was collected. This project uses 29 real GDP growth rates and 196 market share percentages from 1996 to 2023 to compare two time series data sets.

The original plan was to set a time frame from 1995 to 2023 and compare GDP growth rate with the market share percentage. However, it was decided that comparing the changes in popularity of genre to shifts in economic situation would be more appropriate. Therefore, the rate of change in market share percentage was calculated as follows: Year 2 Market Share Percentage - Year 1 Market Share Percentage. As a result, the first year, 1995, was omitted. On the other hand, for analysis of volatility of each genre's popularity and interrelationship between genres, market share percentage was used.

Data Cleaning

Column headers in both data sets were changed to enable comparison and provide clear interpretation of values inside. Then, different data format conversion and additional calculation were applied to each data set as following:

	Original Data Format	Converted Data Format	Changes	Function/Method	Purpose
GDP growth data set	Year: MM/DD/YYYY GDP growth rate: float	Year: YYYY GDP growth rate: float	Added "Year" column	1. Original date data type was converted to datetime object using "pd.to_datetime()" function in Pandas 2. Extracted year parts	To match with the market share data set
Market Share by Genre data set	Year: integer Genre: integer Market Share: float Market Share Delta: NA	Year: integer Genre: string Market Share: float Market Share Delta: float	Converted values in "Year" and "Genre" columns	1. Created two dictionaries,"genre_ID" and "year_ID" 2. Updated original values through ".replace()" function in Pandas library.	To replace the original integer IDs with the interpretable values
			Added "Market Share Delta" column	Rate of change in market share percentage was calculated from the original data using ".diff()" function in Pandas library.	To compare popularity change to economic shifts

The processed data sets were saved as .csv files in the data/processed folder. Lastly, the saved files were merged into a single data frame based on the "Year" column for analysis and visualization purposes, and the resulting data frame was saved in the same folder as "merged.csv." As the two data sets merged and the rate of change in market share was calculated, the time range of the final data set was set to 1996 to 2023.

Data Analysis/Visualization

This project adopted Ordinary Least Squares (OLS) regression, descriptive statistics, and Pearson's correlation coefficient to analyze and find patterns in movie genre popularity trends. Visualizations are generated through seaborn and matplotlib.pyplot libraries.

1. Relationship between economic cycle and genre popularity

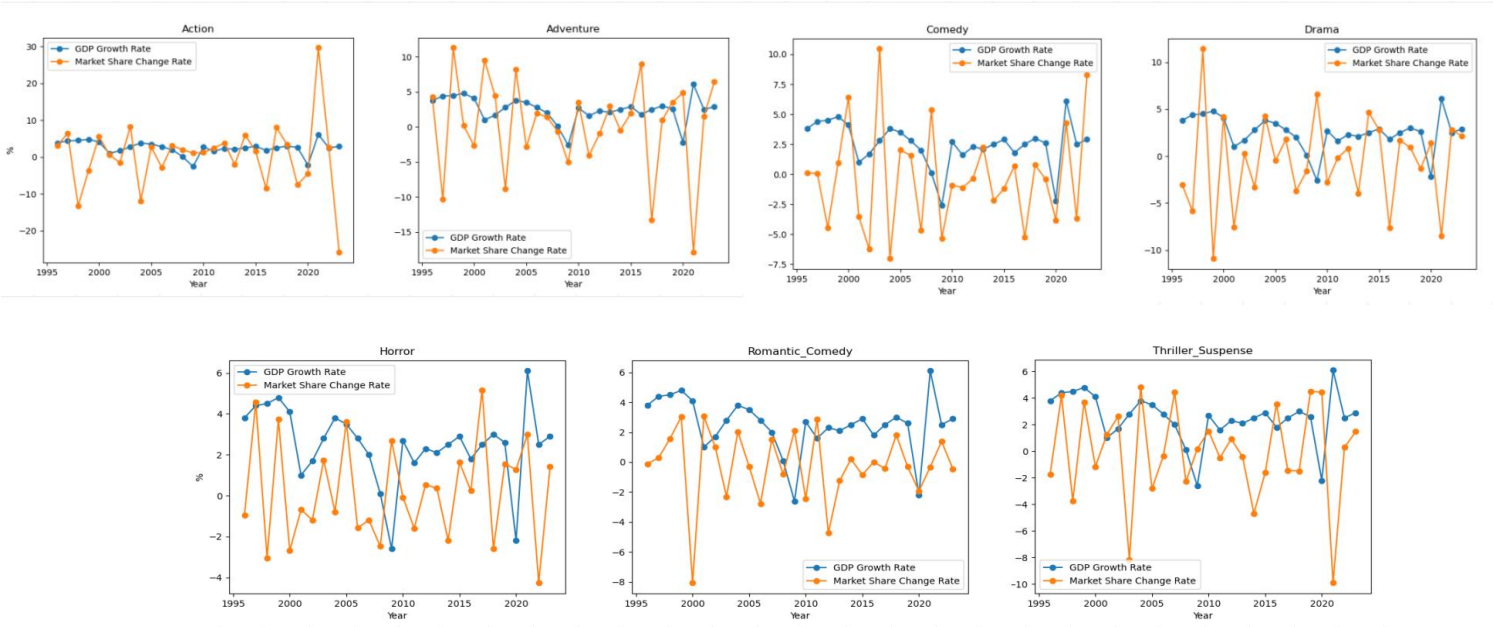
OLS regression analysis through statsmodels.api library was used to study the relationship between the economic cycle and the genre popularity trend. The independent variable is real GDP growth rate and the dependent variable is the rate of change in market share for each genre. The following table summarizes results of key indicators:

	R-Squared	P-Value
Action	0.038	0.320
Adventure	0.039	0.315
Comedy	0.085	0.133
Drama	0.043	0.288
Horror	0.005	0.715
Romantic Comedy	0.007	0.678
Thriller/Suspense	0.088	0.125

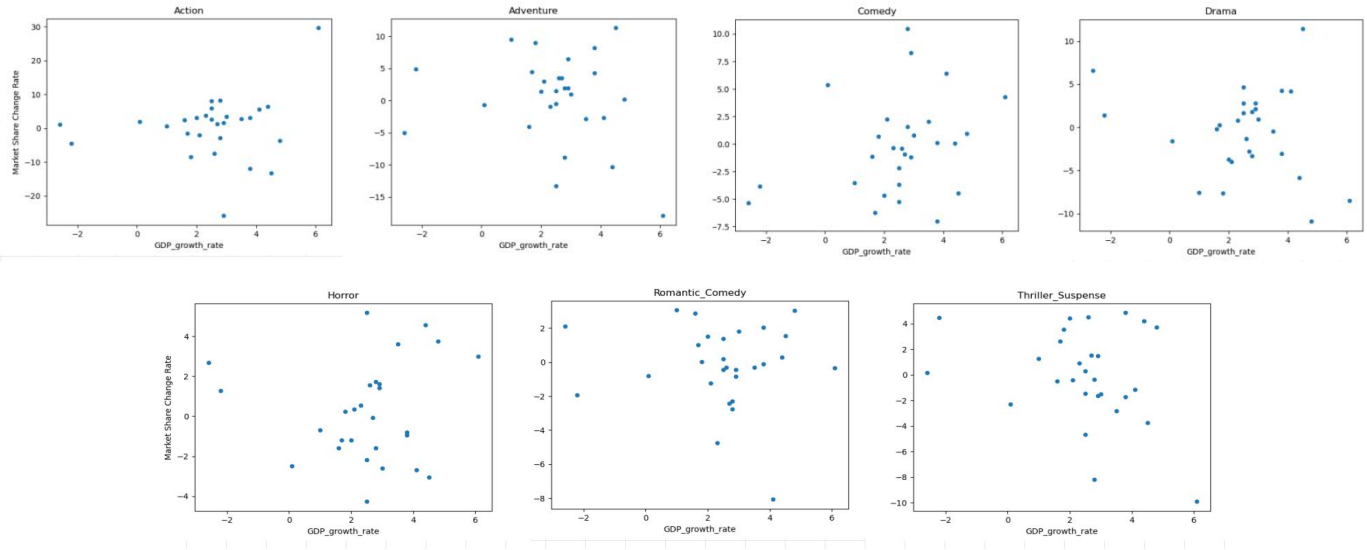
In general, R-squared above 0.7 indicates a strong relationship between the independent dependent variables as the it means that the model accounts for at least 70% of the variance in the dependent variable. However, resulting R-squared values for each genre was below 0.1, indicating the model failed to find a statistically significant relationship between GDP growth rate and rate of change in market share.

The p-value represents the likelihood of observing a test statistic as extreme or more extreme than the one calculated from the data under the assumption that the null hypothesis is true. A p-value greater than 0.05 suggests there is insufficient evidence to establish a significant relationship between the two variables. In this case, the null hypothesis states no relationship exists between real GDP growth rate and rate of change in market share. The calculated p-values exceed 0.05. Therefore, the results indicate there is no strong evidence of a significant correlation.

Below are the visual representations of the results.



The X-axis of the line graph represents time and the Y-axis represents the rate of change in both real GDP and market share. As shown, the rate of change of market share fluctuates more often and a consistent pattern cannot be found when it is compared to the economic cycle.



The X-axis of the scatter plots represents the independent variable used in regression and the Y axis represents the dependent variable. In this case, real GDP growth rate and market share change rate. As shown, the points are scattered randomly, indicating that there is a weak relationship between the variables.

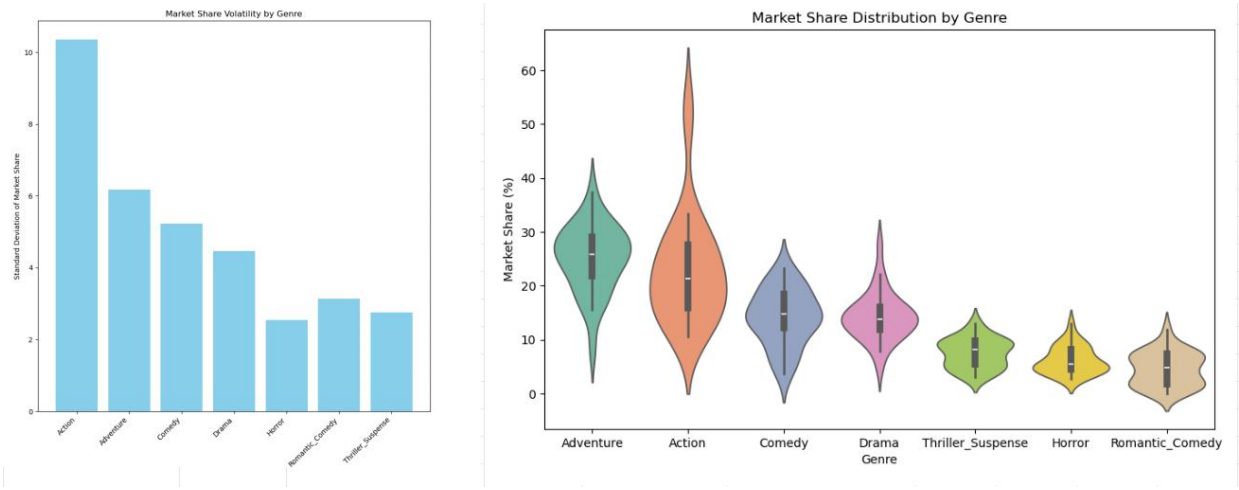
2. Fluctuation/Volatility of each genre's popularity

To examine volatility or fluctuation of the popularity for each genre, standard deviation from descriptive statistics was used through .describe() function.

Genre	count	mean	standard deviation	min	25%	50%	75%	max	range
Action	28	22.99	10.36	10.61	15.92	21.28	27.57	53.56	42.95
Adventure	28	25.11	6.17	8.44	21.78	25.89	29.09	37.36	28.92
Comedy	28	14.22	5.22	3.74	12.26	14.71	18.50	23.29	19.56
Drama	28	14.29	4.46	5.05	11.87	13.89	15.99	27.67	22.63
Horror	28	6.21	2.55	2.69	4.51	5.49	8.22	12.94	10.25
Romantic_Comedy	28	4.58	3.14	0.03	1.74	4.76	7.32	11.91	11.88
Thriller_Suspense	28	7.54	2.76	3.12	5.44	8.15	9.76	13.03	9.91

As shown in the third column, the standard deviation of the action genre's market share is the highest, meaning that the popularity fluctuates the most over time. However, the action genre also has comparably high average market share. On the other hand, mean value and standard deviation of horror, thriller/suspense, and romantic comedy genres' market share are comparably low, indicating that it has small but relatively stable audiences.

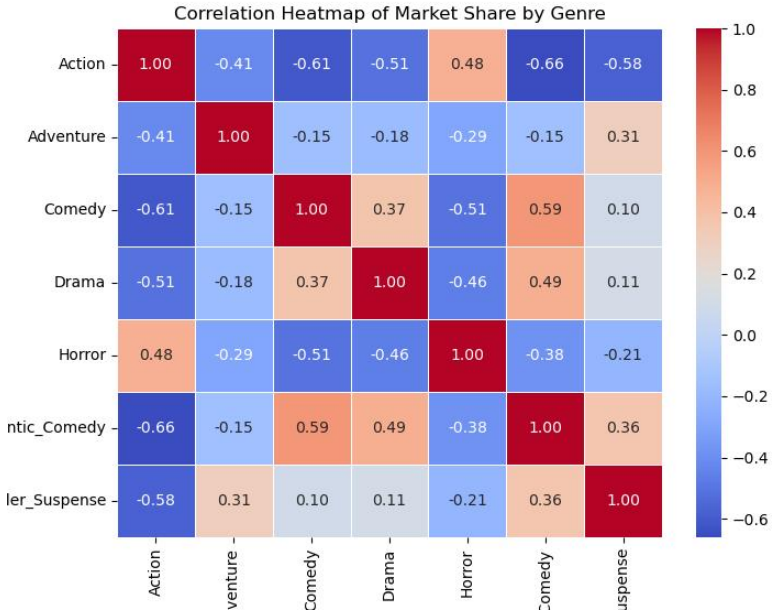
Below is the visual representation of market share volatility comparison and distribution.



The violin plot on the right provides a general idea of market share distribution. It represents the density of the data points at a given value, meaning the length of the plot corresponds to the range of data. In this case, length of the action genre is the longest, which is consistent with the result from descriptive statistics calculation and the bar graph on the left; it has highest volatility.

3. Interrelationship between genres.

Lastly Pearson's correlation coefficients are calculated by ".corr()" function to see the interrelationship between genres. It is represented through the heat map.



Negative values indicate the two genres are inversely related. In this case, action and romantic comedy got -0.66, showing that as market share of action movies increases, market share of romantic comedies tend to decrease. On the other hand, romantic comedy and comedy have a strong positive correlation, implying that these genres often perform well together. Many genre pairs, such as Drama and Thriller/Suspense (0.11) or Comedy and Thriller/Suspense (0.10), show very weak correlations, suggesting that their market performances may be independent.

Conclusion

In conclusion, the analysis of movie genre market share dynamics revealed several important findings. First, no strong evidence was found to support a significant correlation between GDP growth rate and the change in market share data, suggesting that economic performance may not directly influence the shifts in movie genre popularity. Second, the analysis of market share volatility using descriptive statistics highlighted that the action genre experiences the most fluctuations over time, indicating that its popularity varies significantly. However, it should be noted that it also has a relatively high average market share. In contrast, genres such as horror, thriller/suspense, and romantic comedy have lower mean values and standard deviations, pointing to more stable but smaller audiences. Lastly, the correlation analysis result suggests that there is an inverse relationship between action and romantic comedy genres while there is a positive relationship between romantic comedy and comedy genres. However, many other genre pairs showed weak correlations, suggesting that market performances may be largely independent of one another.

Future Work

There are several ways to improve this study. Using alternative economic indicators, such as the unemployment rate, economic misery index, or interest rates, instead of the real GDP growth rate, may reveal more substantial relationships. Additionally, to measure the popularity of certain genres, other indicators, such as tickets sold or revenue, could be utilized. Furthermore, data sets can be separated into two groups: pre-COVID period and post-COVID period, to remove outliers and minimize bias created by the the special event that affect the industry.

I would continue to use the FRED API to gather economic performance indicators as it provides various time series data, including GDP, consumer price index, and unemployment rates. However, to measure a certain genre's popularity, I would consider using a different source, as the current The Numbers web page only provides the calculated market share percentage for each genre.

Reference

Lee. (2019). " A Study on the Relationship between Korean Economic Indicators and Genres of the Box Office Trend of Movies and Pop Music." *Chung-Ang University Graduate School of Art*. doi: 10.23169/cau.000000230461.11052.0000475