**Analysis of the Correlation of Movie Ratings on Movie Revenue.**

Results of Analysis

**A: Project Highlights**

What correlation exists between a movie’s revenue and user ratings? Do more highly rated movies make more money? This analysis aims to identify if such a correlation exists and to what extent it affects revenue.

This project’s scope includes gathering, cleaning, and exploring two source datasets to identify relationships between movie revenue and numerical user ratings. The method used to identify correlation will be the Scipy Stats library, which will identify any correlation and probability for each year over a ten-year period from 2006 to 2016. Given the results over this period, the range was also expanded to a thirty-year period, from 1986 to 2016, with similar results.

This project will not include analysis of other impact factors such as professional critic ratings, movie genre, box office distribution, or any other unidentified factor. The datasets initially selected for this analysis shall not be expanded upon after the commencement of this project. Any further analysis identified herein should be pursued in a follow-up project.

The tools used to perform this analysis included Pycharm and Jupyter Notebooks using Python code. Libraries to be utilized in this analysis shall include Pandas, Scipy, Matplotlib, datetime, JSON, requests, and os. Additional libraries may be included, as necessary.

**B: Project Execution**

The goal is to identify if any statistical significance exists between user ratings and movie revenue.

The project objective is to determine the correlation and probability between a movie’s user ratings and revenue and identify any year-over-year trends.

Deliverables include the specific correlation and P value for movie revenue vs. user ratings for each year in the sample over the initial sample period, 2006 to 2016. Deliverables include a scatterplot and two bar charts for each year in the sample. The scatterplot compares revenue and average user rating for each movie release during the given year. For the bar charts per year, one chart lists the top ten movies by total revenue, and the other lists the top ten movies by top user ratings. Both charts include the percentage ranking of each of the included movie’s total revenue and user rating.

The correlation values and P values numerically identify the correlation and likelihood for our hypothesis while the various charts visually demonstrate the discovered relationships.

The methodology for the implementation of this project is ADDIE.

* The analysis phase included the gathering and exploration of the selected datasets for this project.
* The Design phase included the identification of required data and planning of the steps necessary to convert the original dataset(s) into useful data. This phase also included the identification of the hypothesis and null hypothesis for this project.
* The Development phase includes the required data cleaning, data conversion, and joining of datasets to create a single dataset for analysis, as well as the writing of the code necessary to implement the methods to assess the correlation coefficients, P values, and sample size data for each year’s movies.
* The Implementation phase included the implementation of created methods to capture the data required for this project. This phase will also include the creation of the scatterplots and bar charts per year.
* The evaluation phase included the review of analysis results to determine whether the hypothesis or null hypothesis was found correct.

Project Timeline:

| **Milestone or Deliverable** | **Duration** | **Start Date** | **End date** |
| --- | --- | --- | --- |
| Cleaned, joined dataset ready for analysis | 2 days | 12/1/2024 | 12/3/2024 |
| Availability of classes and methods to assess correlation coefficients, P\_values, sample data per year, and related plots. | 2 days | 12/3/2024 | 12/5/2024 |
| Availability of analysis results and required plots. | 1 day | 12/5/2024 | 12/6/2024 |
| Analysis of project results. | 2 days | 12/6/2024 | 12/8/2024 |

**C: Data Collection Process**

Two source datasets were found to facilitate the analysis of relationships between movie revenue and user ratings. These are the same datasets identified in the project proposal.

1. The Revenue dataset:

From OpenML: Detailed movie descriptions - ideal for Recommendation Engines Website: <https://www.openml.org/search?type=data&status=active&id=43113> Link to data: <https://www.openml.org/data/download/22047889/dataset>

The Revenue dataset will be collected programmatically using the download link: <https://www.openml.org/data/download/22047889/dataset> and stored locally.

1. The Review dataset:

From OpenML: This dataset contains IMDb ratings and votes information for movies having original titles. Website: <https://www.openml.org/search?type=data&sort=runs&id=43784&status=active> Link to data: <https://www.openml.org/data/download/22102609/dataset>

The Review dataset has been downloaded using the OpenML download link: <https://www.openml.org/data/download/22102609/dataset> and stored locally.

One obstacle was found during the data-gathering process. During the import of the source data into working data frames, errors were found when importing a small portion of the total rows from the source CSV files. These include 67 rows from the Revenue dataset which included 4715 rows originally which comes out to 1.4% of the source dataset not being included. Similarly, the Review data frame, which contained 67408 rows of data, included 1824 rows of data that could not be imported from the CSV file which comes out to 2.7%. These rows of data were not included in the analysis and a ‘bad\_line\_logger’ function was created to capture for future reference if necessary. The remaining data was determined to be sufficient for this examination of the relationship between movie revenue and user ratings. No other obstacles were encountered.

Regarding unplanned data governance issues, given the project’s scope, public domain source datasets, and lack of confidential information, no unexpected issues were encountered.

**C1: Advantages and Limitations of Data Set**

An advantage of the source datasets is their small size specifically related to the number of columns in the datasets. The Review dataset contains only 5 columns of data which keeps the cleaning of the data frame to a minimum regarding data types and removal of unnecessary data. The Revenue data frame contains 10 columns of data which remains very manageable. While there are other datasets available from which to perform this examination, those datasets include much data that is not useful to the analysis given the specific scope of this project.

One disadvantage of these datasets is the lack of additional columns which, while not useful per this project’s scope, could provide insights during this analysis that could lead researchers to other more relevant factors related to movie revenue to investigate in future projects. In this case, the small size of the source datasets could be both an advantage and a potential disadvantage.

Another disadvantage of the source datasets is the lack of specific sources of a movie’s revenue such as box office opening revenue, revenue from streaming services, etc. The breakdown of the total revenue into these sources could provide greater insight into the relationship of user ratings related to these specific sources of a movie’s revenue.

**D: Data Extraction and Preparation**

Data preparation was performed as anticipated with little deviation from the cleaning steps identified in the project proposal, except for the import from CSV process.

During the import of the source data into working data frames, errors were found when importing a small portion of the total rows from the source CSV files. These include 67 rows from the Revenue dataset which included 4715 rows originally which comes out to 1.4% of the source dataset not being included. Similarly, the Review data frame, which contained 67408 rows of data, included 1824 rows of data that could not be imported from the CSV file which comes out to 2.7%. These rows of data were not included in the analysis and a ‘bad\_line\_logger’ function was created to capture for future reference if necessary. The remaining data was determined to be sufficient for this examination of the relationship between movie revenue and user ratings. The remaining processing steps were performed with no unexpected issues.

These datasets contained more information than necessary to prove the hypothesis regarding movie revenue and ratings. The only needed columns were the “id,” “title,” “original\_title,” “revenue”, “review”, and “averageRating”. The inclusion of the remaining fields could negatively impact the readability of the cleaned and joined datasets so these unnecessary columns were removed.

The data types for the “revenue” and “averageRating” columns were imported as strings; however, these fields needed to be a “float”. Due to inconsistency in the source data, I had to import all fields as strings. The data types for the “revenue” and “averageRating” columns were converted into floats using the “.astype()” function.

The Revenue dataset had 1396 rows with '0' revenue listed. As these rows do not contain information useful to proving the hypothesis, these rows were remove.

The Review dataset contained many average ratings based on relatively few reviews. The minimum number of user ratings required to include the average user rating in this analysis has been set to five hundred. Any rows with less than five hundred reviews contributing to the average rating have been removed.

In the Revenue dataset, there were two columns containing movie title data: the “title” column and the “original\_title” column. The "original\_title" column has been removed as it contains alternate titles, includes special characters, and does not match the title data in the Review dataset. The movie Revenue dataset’s ‘title” column does match the Review dataset’s “title” column and was useful in this analysis.

**E: Data Analysis Process**

**E1 Data Analysis Methods**

The P-value and correlation coefficients were identified using “scipy.stats.pearsonr” comparing movie revenue and user ratings. A function was developed to filter the cleaned and joined datasets by year. SciPy’s Pearsonr function was then applied to this filtered dataset to calculate the correlation coefficient and p\_value for the year. Code was also implemented in this function to create a scatterplot and two bar charts to visually demonstrate any relationship. The bar charts each list the top ten movies with one chart sorted by user ratings and the other by revenue. Both charts show the total revenue and user ratings per movie. A call applying this function over a range of ten years then provided the data required for evaluation of the null hypothesis.

Given the size and content of the joined working data frame, the Pearsonr function is appropriate for calculating the correlation coefficient and P value for each year in the sampled year range; 2006 to 2016.

**E2 Advantages and Limitations**

One advantage of this Scipy Stats Pearsonr is its ease of use. This function requires only the source data series and returns the correlation coefficient and P value.

One limitation of this tool is that it is limited to extrapolation of linear relationships. Should the data have been more complex, being limited to a linear relationship may not have been appropriate. For the needs of this analysis, Pearsonr is ideal.

**E3 Application of Analytical Methods**

A function was developed to filter the cleaned and joined datasets by year. SciPy’s Pearsonr function was then applied to this filtered dataset to calculate the correlation coefficient and p\_value for a given year.

To visually demonstrate any relationship, this function also contains code to create a scatterplot and two bar charts. The bar charts each list the top ten movies with one chart sorted by user ratings and the other by revenue. Both charts show the total revenue and user ratings per movie. A call applying this function over a range of ten years then provided the data required for evaluation of the null hypothesis.

The correlation coefficient, P value, and sample size for each year in the sampled ten-year period from 2006 to 2016 are also recorded and plotted to identify any trends in these features over the sample period.

To further identify trends in these features one additional run was performed including data over a thirty-year period from 1986 to 2016.

Verification of the functions was performed over several iterations during the development of the code until repeatable results were achieved.

**F Data Analysis Results**

**F1 Statistical Significance (or Model)**