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

**A: Proposal Overview**

**A1: Research Question**

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.

**A2: Context & Background**

Many people believe that a movie is “good” based on reviews and ratings it receives from other viewers and that “good” movies will earn more revenue than “bad” ones. This could drive potential movie viewers to only watch movies that receive high ratings, driving producers to create content based on prior examples of “good” well-rated movies. This analysis will compare movie ratings to their revenue to identify if user ratings are a good measure of a movie’s potential success.

**A3: Summary of Works & Relation to Project**

1. **The Impact of Movie Reviews on Box Office: Media Portfolios and the Intermediation of Genre**

In this study, Koschat examines the performance of 750 titles released between 2004 and 2008 to identify the effect of professional reviewers’ ratings on box office revenues. While it is discussed that the three most important factors are the opening theater count, viewer sentiment, and the word-of-mouth effect, the focus of this study is the impact of professional reviewers’ ratings.

Koschat identifies a high correlation between viewers’ and (professional) reviewers’ ratings. This study determined that positive reviewer ratings had no discernible impact on the opening weekend box office for non-literary genres; however, negative reviewer ratings “are indicative of low viewer ratings, which do have a demonstrably negative and significant economic impact.” It is also noted that reviewers' ratings do “not affect revenues of literary (and other) titles for the remainder of the release period” as reviews distributed in traditional media are only available for a limited time. (Koschat, 2012)

Koschat identified that reviewer ratings may have a significant impact on box office releases. We intend to explore this impact further.

1. **Investigating the effects of textual reviews from consumers and critics on movie sales**

In this paper, Deng investigates “the sales impact of different types of online word-of-mouth based on their source (user vs critic) and form (structured vs unstructured).” In this study, both unstructured textual data and numeric ratings from critics and users are examined. Two streams of word-of-mouth are researched: discrepancies in online word-of-mouth, and the sales impacts of user and critic reviews. (“Investigating the Effects of Textual Reviews From Consumers and Critics on Movie Sales,” 2019b)

This study resulted in two notable findings: there is a demonstrable discrepancy between user and critic reviews, and of the four types of reviews, two have significant impacts. Critics’ textual reviews and users’ numeric ratings are identified as the sources of reviews that impact purchase decisions.

Regarding this project, we will focus on the relationship between users’ numeric ratings and movie revenue. Per Deng, users’ numerical ratings may have a significant impact on movie revenue. We will explore this further.

1. **Is everybody an expert? An investigation into the impact of professional versus user reviews on movie revenues**

In this study, electronic word-of-mouth (eWOM), in the form of user and expert reviews, is examined to determine the impact thereof on movie revenues. A random sample of 194 films from 2007 to early 2008 were selected for this study. User and expert reviews for the first ten weeks following the film’s release were studied against weekly box office revenue to identify any significant impact. Other variables are identified such as genre, MPAA ratings, and star power; however, the focus of this study is the impact of user and critic reviews.

It is determined that “the impact of critic valence on revenues is significantly larger than the corresponding effect of user valence.” (Basuroy et al., 2019) They found that word-of-mouth does not seem to impact moviegoing, while it is significantly impacted by expert reviews. They also found that “when the moviegoing public feels very differently than the experts, then user ratings start to matter more.” (Basuroy et al., 2019)

This study concludes that expert reviews matter more than user reviews relating to movie revenues and that user ratings have no impact on wide-release movies, except in the case where there is a large discrepancy between expert and user reviews; however, user ratings remain less impactful than expert reviews in determining a movie’s success.

We intend to identify if a relationship exists between user ratings and movie revenue over a larger sample of movies. Based on the results of Basuroy et al’s study above, it is likely that user reviews will have no significant impact; however, the results may be different given the increased sample size and revenue over time for the sample we have selected.

**A4: Summary of Data Analytics Solution**

In this project, the intention is to select a sample of movie ratings and revenues over a given period of years and identify if any correlation exists between the movie’s overall user ratings and total revenue. The sampled movies will be grouped by year for this examination and the identified correlation per year will be compared to identify trends.

Correlation will be determined using linear regression between the movie revenue and user rating of the sample set. The dataset from which our yearly samples will be pulled includes various features, including revenue and ratings, for films from 1925 to 2016. We will begin with the ten years from 2006 to 2016, and based upon the results we may broaden the sampled years.

The original dataset(s) from which the data will be analyzed will be available online. The intention is to either have the data already saved locally or to use an API to get the data and store it locally. The dataset(s) will be public domain and available online from a repository such as OpenML or an equivalent public repository.

To get the data we need for this project, multiple datasets may be necessary. The original datasets will need to be cleaned, joined, and filtered. Cleaning will include removing unnecessary features and dropping instances with missing data. Then, after ensuring the remaining data is in the correct format, we will run methods to filter the data by year and capture movie ratings, movie revenue, and the sample sizes for each year included in the run.

Once the regression is complete, we will demonstrate the relationship visually using scatterplots. To further demonstrate any relationship, or lack thereof, between movie revenue and user ratings, we will also include two additional bar charts comparing these results for the top ten movies, listed based on highest revenue and by highest user rating, from each year sampled. These bar charts will show the ratings for the top ten movies based on revenue and then the revenue for the top ten movies based on user ratings. Any correlation should be easily identifiable in these three charts: the scatterplot and the two bar charts.

**A5: Benefits & Support of Decision-Making Process**

The results of this project will identify the relationship between movie revenue and user ratings. Understanding this relationship will help users identify the value of user ratings in their decision-making process. This understanding will also be useful to movie creators in determining key indicators for the success of their content, such as properly assigning value to user ratings when deciding on what content to create.

Should correlation be found statistically significant, more value should be given to user ratings in decision-making; however, should no correlation be found, additional factors may need to be explored to determine indicators of a movie’s potential.

**B: Data Analytics Project Plan**

**B1: Goals, Objectives, & Deliverables**

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. Deliverables shall also include a scatterplot and two bar charts for each year in the sample. The scatterplot shall compare a movie’s revenue and user rating. For the bar charts per year, one chart shall list the top ten movies by total revenue, and one shall list the top ten movies by top user ratings. Both charts shall include the percentage ranking of each of the included movie’s total revenue and user rating.

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

**B2: Scope of Project**

This project will include gathering, cleaning, and exploring the source dataset(s) to identify relationships between movie revenue and numerical user ratings. The methods used to identify correlation and the years to be sampled may be expanded based on the results of analysis for the ten-year range from 2006 to 2016.

This project will not include an analysis of the impact of other 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.

**B3: Standard Methodology**

The methodology for the implementation of this project shall be ADDIE.

* The analysis phase shall include the gathering and exploration of the selected datasets for this project.
* The Design phase shall include the identification of required data and planning the steps required to convert the original dataset(s) into useful data. This phase shall also include the identification of the hypothesis and null hypothesis for this project.
* The Development phase shall include the required data cleaning, data conversion, and joining of datasets to create a single dataset for analysis. This phase shall also include writing any required code to implement methods to assess the correlation coefficients, P values, and sample size data for each year’s movies.
* The Implementation phase shall include the implementation of the 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 shall include a review of analysis results to determine whether the hypothesis or null hypothesis has been found correct.

**B4: Timeline & Milestones**

| **Milestone or Deliverable** | **Duration** | **Projected Start Date** | **Projected 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 |

**B5: Resources & Costs**

1. Hardware (Computer): No cost (utilize existing hardware)
2. Software (Source Data): No cost (public domain dataset(s))
3. Software (Pycharm): No cost (utilize existing / free software)
4. Software (Jupyter Notebook): No cost (utilize existing / free software)
5. Thirty-five work hours:  $1750 (35 hours at $50 per hour)

**B6: Criteria for Success**

Availability of results of the analysis to include all deliverables as per section B1 above:

* The specific correlation and P value for movie revenue vs. user ratings for each year in the sample.
* A scatterplot and two bar charts for each year in the sample. The scatterplot shall include percentage ranking for each movie’s total revenue and user rating. For the bar charts per year, one chart shall list the top ten movies by total revenue, and one shall list the top ten movies by top user ratings. Both charts shall include the percentage ranking of each of the included movie’s total revenue and user rating.

**C: Design of Data Analytics Solution**

**C1: Hypothesis**

Alternate Hypothesis: Higher user ratings correlate with higher movie revenue.

Test: correlation coefficient >= 0.7, p\_value =< 0.05

Null Hypothesis: Higher user ratings do not correlate with high movie revenue.

Test: correlation coefficient =< 0.7, p\_value > 0.05

**C2: Analytical Method**

A correlation coefficient >= 0.7 will indicate that the alternate hypothesis is statistically significant.

A p\_value =< 0.05 will demonstrate that the alternate hypothesis is highly likely to occur.

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

**C3: Tools & Environment**

This analysis will be completed using Pycharm and Jupyter Notebooks and 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.

**C4: Methods & Metrics to Evaluate Statistical Significance**

Null Hypothesis: Higher user ratings do not correlate with high movie revenue.

Alternate Hypothesis: Higher user ratings correlate with high movie revenue.

The test will assess the Pearson Correlation Coefficient and P-value between a movie’s total revenue and user rating using the scipy.stats library. This method is appropriate for this study given the size of the datasets and the project scope.

The Alpha value will be set at 0.05.

If the P-value is less than 0.05, the null hypothesis will be rejected and the alternate hypothesis accepted.

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

**C5: Practical Significance**

The practical significance of the results of this project lies in the proper assignment of value we each give to a movie’s user ratings. Many individuals base their purchases on the results of peer review, and this basis for determining the value of a product to you may be flawed. Each of us assesses a product based on another user’s ratings to some extent prior to purchasing that product. The goal of this project is to demonstrate if numerical user ratings are useful in determining the success of a product which in this case, the products are box office movies.

**C6: Visual Communication**

Graphic visualizations of the results of this project shall include one scatterplot and two bar charts. One set of these visualizations shall be created for each of the ten years in the sampled movie data.

The scatterplot shall compare the movie revenue and user ratings.

Bar chart 1 shall list the top ten movies of the given year by revenue. The listings will be as a percentage of the highest revenue generated for any movie in the entire dataset. This bar chart will include two bars: one for the percent total revenue and one for the percent of highest user rating.

Bar chart 2 shall list the top ten movies of the given year by user rating. The listings will be as a percentage of the highest rating generated for any movie in the entire dataset. This bar chart will include two bars: one for the percent total revenue and one for the percent of highest user rating.

**D: Description of Dataset**

**D1: Source of Data**

The two datasets selected for this project are public domain and are found on OpenML.org.

The two datasets are:

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>

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>

**D2: Appropriateness of Dataset**

Both datasets identified above provide some of the data required to compare movie revenue to user ratings. The Revenue dataset provides movie details including movie titles, release dates, and revenue; however, this dataset does not include user ratings. The Review dataset includes movie titles and details the average user rating and number of user votes. Once cleaned and combined the resulting dataset will provide all the data required to perform this analysis.

**D3: Data Collection Methods**

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

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

**D4: Observations on Quality & Completeness of Data**

These datasets contain more information than necessary to prove the hypothesis regarding movie revenue and ratings. The only needed columns are the “id,” “title,” “original\_title,” “revenue”, “review”, and “averageRating”. The inclusion of the remaining fields could negatively impact the readability of the cleaned datasets, especially after they are joined.

The data types for the “revenue” and “averageRating” columns should be a float. To get the data uploaded I had to do all fields as strings due to inconsistency in the source data. There is data in this column that is text and cannot be converted to a float. This will need to be corrected.

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

The Review dataset contains many average ratings based on relatively few reviews. I will arbitrarily pick a number of reviews to set as the minimum qty and remove any rows with less than this qty of reviews contributing to the average rating.

In the Revenue dataset, there are two columns containing movie title data: the “title” column and the “original\_title” column. The "original\_title" column will be 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.

**D5: Data Governance, Privacy, Security, Ethical, Legal, & Regulatory Compliances**

Data governance is a part of every project and requires consideration of accountability, integrity, standardization, data quality, privacy, security, etc. The implications of these considerations on this project are small, given the project’s scope and public domain source data. The accountability for this project falls on me as this is my project. The data selected was specifically chosen as it is public domain and free of sensitive data. The project data will be stored locally and in a public Git repository to be shared with WGU reviewers and as such has the minimum data security concerns.

Data Privacy is not required for this project as the data included in this project is public domain, available on OpenML, and does not contain any private or personally identifiable information. Should any confidential data have been included in this project, measures would have been necessary to protect that data from unintended use.

Security is a minimal concern for this project as it utilizes information that is public domain and does not contain confidential data or PII. This project is stored locally on a personal computer with controlled access and in a public Git repository. Read access to this project is available to anyone who can find it. Should this data contain sensitive information it would be necessary to ensure the data remain secure, with access controlled.

Ethical, legal, and regulatory compliance considerations include ensuring all sources of information are properly credited and referenced. Other legal and regulatory compliance is not applicable to this project due to the scope and nature of the analysis being performed. Should sensitive or controlled information have been used in this project, it would be necessary to be aware of any related legal or regulatory implications.

**References:**

* *Koschat, M. A. (2012). The impact of movie reviews on box office: media portfolios and the intermediation of genre. Journal of Media Economics, 25(1), 35–53.* [*https://eds.p.ebscohost.com/eds/detail/detail?vid=0&sid=d931725b-0b04-41cc-bbd4-bc9d748fdbad%40redis&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=73325511&db=buh*](https://eds.p.ebscohost.com/eds/detail/detail?vid=0&sid=d931725b-0b04-41cc-bbd4-bc9d748fdbad%40redis&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=73325511&db=buh)
* Deng, T. (2020). Investigating the effects of textual reviews from consumers and critics on movie sales. *Online Information Review*, *44*(6), 1245–1265. <https://www.proquest.com/docview/2453640662?accountid=42542&sourcetype=Scholarly%20Journals>
* Basuroy, S., Ravid, S. A., Gretz, R. T., & Allen, B. J. (2019). Is everybody an expert? An investigation into the impact of professional versus user reviews on movie revenues. Journal of Cultural Economics, 44(1), 57–96. [Is everybody an expert? An investigation into the impact of professional versus user reviews on movie revenues - ProQuest](https://www.proquest.com/docview/2220517023?accountid=42542&sourcetype=Scholarly%20Journals)