# Box Office Analysis

***Introduction:***

This project aims to comprehensively analyze the impact of online consumer word-of-mouth, specifically movie reviews, on box office performance. The primary focus is understanding how review features like overall rating and review count influence box office success.

Additionally, we explore various other variables potentially affecting box office performance, including genres, movie length, ticket price, trailers, pictures, awards, and number of screens. Utilizing movie review data, we performed various analysis, such as correlation, scatter plots for dependent and independent variables.

After analyzing all data science techniques, we interpreted the different independent variables' impact on the dependent variable (box office performance).

***Data Processing****:*

For our project, we were given two data sets box\_office\_data and movie\_review\_data which contained common variables movieID and date. Assuming a one-day delay in the date we added a column named next\_date to the sheet using the formula date+1. Later, we merged the two datasets using VLOOKUP by making use of common variables. The merging resulted in 2810 rows and 18 columns.

***Data exploration:***

This merged dataset contained nonnumeric values like ‘#N/A’ in box\_office column and there were empty cells in columns IMDb, genre2 and genre3 which holds further analysis. IMDb is very unlikely to affect the box\_office the column is dropped.

***Dummy Variable – “Award”:***

The ‘award’ column consists of only ‘yes’ or ‘no’ values which can be converted into numeric values by assigning 'yes’ to 1 and no to 0. This way categorical variables are converted into numerical form

***Analyzing the Variables:***

Variables are categorized into independent and dependent types. The box office performance acts as the dependent variable, influenced by several independent variables, which can be further divided into word-of-mouth and non-word-of-mouth factors. These independent variables collectively shape the box office performance, which serves as the dependent variable.

Typically, **Box\_office** is selected as dependent variable for our project which analyses the box office performance based on other independent variables as follows.

**Independent variables:**

**Word-of-mouth**

1. Overall rating
2. Movie reviews

**Non-word-of-mouth:**

1. Movie genre
2. Ticket price
3. Release date
4. Screens
5. Trailer
6. Pictures
7. Award
8. Length of the movie

***Expectation of relationship between independent variable and dependent variable:***

The connection between variables and a movie's box office success can be categorized as positive or negative. Here are some anticipated effects of various elements on a film's box office performance.

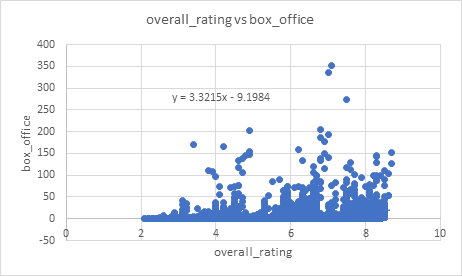
* **Overall rating:**
  + A higher overall rating is likely to result in better box office performance, as it indicates that audiences have reacted positively to the movie, making them more inclined to watch it, thus driving up ticket sales.
* **Reviews:**
  + A high number of reviews indicates higher engagement of audience irrespective of positive, negative or neutral review. Thus, affecting the box\_office positively.
* **Trailer:**
  + Audiences anticipate an enjoyable experience when a captivating trailer is released. Compelling visuals and engaging previews generate excitement, heightening expectations for the upcoming film or show. A sturdy trailer for a movie is likely to draw in a higher number of viewers. thereby having a positive effect on the performance of the box office.

* **Pictures:**
  + A positive relationship is expected. The count and quality of the visually striking images/posters create anticipation and excitement among viewers, leading them to believe that the movie promises an engaging and captivating experience, ultimately contributing to its commercial success at the box office.
* **Award:**
  + A film's likelihood of success is predicted to increase upon receiving nominations or wins. The movie gains greater popularity as a result of this recognition, which may increase box-office sales indicating positive relation.
* **Length:**
  + The connection between a movie's length and its box office performance is unclear. While shorter films may have more screening times than longer ones, the ultimate success depends on how captivating and engaging the movie is for audiences. The length alone does not provide a definitive indication of a film's commercial prospects.
* **Ticket:**
* Expensive ticket prices may have an adverse effect on box office revenue. Staggering ticket costs have the ability to deter moviegoers from visiting theaters, which could lead to a reduction in box office earnings for certain releases.
* **Screens:**
  + A film that is shown in more theaters may draw a larger crowd and earn more money at the box office. The number of screens a movie is exhibited on is expected to positively correlate with its box office performance.
* **Drama:**
  + Dramas tend to resonate with a wider audience because of their relatable and grounded storytelling, which often translates into strong box office performance.
* **Comedy:**
  + Comedies tend to be well-received by audiences due to their entertaining nature and ability to evoke laughter and enjoyment, fostering a positive connection between the film and viewers.
* **Action:**
  + Expectation: A good performance at the box office.
  + Reason: Action movies are popular for their exciting and visually impressive action sequences. Positive reviews and word-of-mouth about thrilling action scenes are likely to draw in audiences looking for an adrenaline rush, leading to better box office results.
* **Animation:**
  + Animation films, with their broad appeal to children and adults alike, tend to foster positive expectations. Favorable reviews and family-friendly content are anticipated to have a positive correlation with box office success, as families are inclined to choose these shared cinematic experiences.
* **Fantasy:**
  + Audiences tend to enjoy fantasy films as they provide an escape from reality by transporting them to imaginative, fictional worlds, which often leads to strong box office performance.
* **Children:**
  + Analysts predict a successful box office run for the latest children's movie release. These films, crafted with young audiences in mind, tend to draw in families seeking entertainment suitable for all ages. Glowing reviews, enthusiastic word-of-mouth, and family-friendly content are expected to drive strong ticket sales as parents opt for these offerings as a shared outing with their kids.
* **Romance:**
  + Expectation: A neutral relationship between the characters is anticipated.
  + Reason: The nature of the relationship is contingent upon the intended audience and narrative arc. Films with romantic themes may cater to a specific niche, potentially influencing box office performance to varying degrees.
* **Thriller:**
  + Thrillers, with their suspenseful storylines, have the potential to create a buzz and generate positive word-of-mouth if they are well-crafted and engaging for audiences.

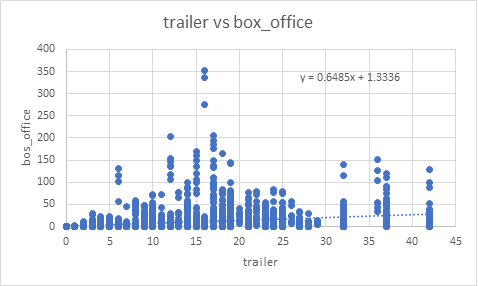
***Data explorations – Scatter plots and Correlation Analysis:***

We conducted data explorations to analyse the relationships between dependent and independent variables.

* **Scatter plot between Overall rating and Box office:**

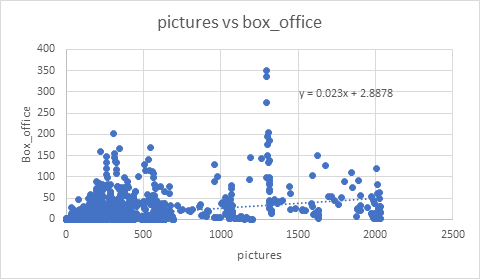


* This scatterplot indicates a positive relationship between a film's overall rating and its box office results. This scatter plot has a larger range of data points than the Screens vs Box Office plot, showing that critical reception may be a less significant predictor of box office success than the number of screens a film is presented on.
* **Scatter plot between trailer and Box office performance:**



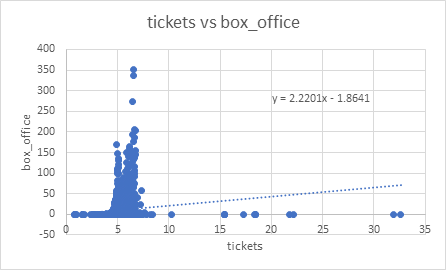
This scatter plot depicts the relationship between the number of trailers and Box Office performance. The positive trend line suggests a positive correlation, indicating that a higher number of trailers is associated with stronger Box Office performance.

* **Scatter plot between pictures and box office performance:**



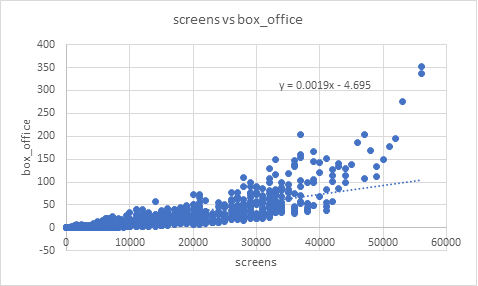
This scatter plot shows the relationship between the number of trailers and Box Office performance. The positive trend line suggests a positive correlation, indicating that a higher number of trailers is associated with stronger Box Office performance.

* **Scatter plot between ticket and box office performance:**



This scatter plot shows the relationship between the number of tickets and Box Office performance. The positive trend line suggests a positive correlation, indicating that a higher number of trailers is associated with stronger Box Office performance. Also majority of the the box\_office sales are concentrated around the ticket prices between 5 and 7.5 implying that higher prices may lead to decline of box\_office sales.

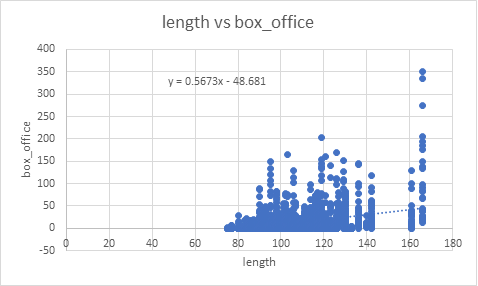
* **Scatter plot between Screens and box office performance**:



This scatter plot depicts the relationship between the number of screens and Box Office performance. The positive trend line suggests a positive correlation, indicating that a higher number of screens is associated with stronger Box Office performance.

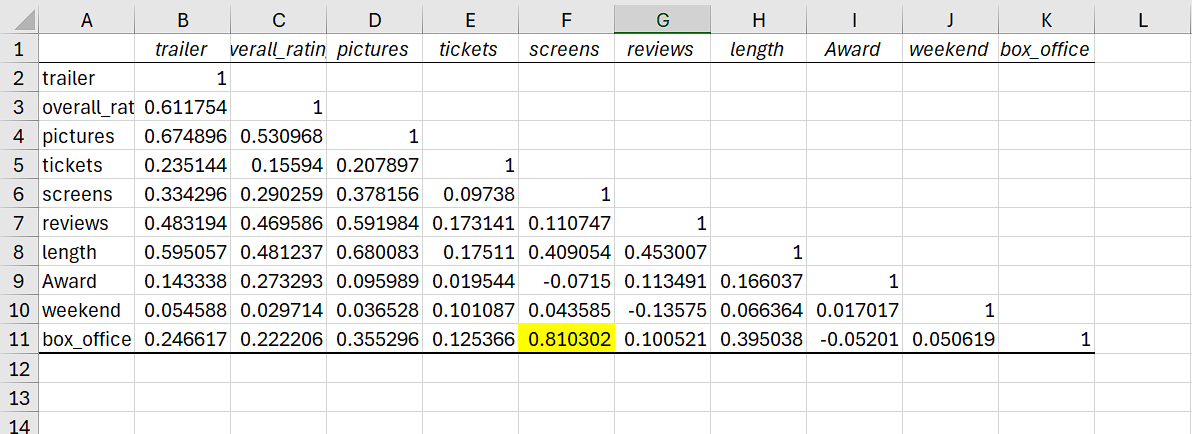
When comparing with box\_office there is an upward trend implying that higher number of screens increase box\_office sales as expected. There are a few data points that fall below the trend line, indicating that some movies may not have performed as well as expected despite being shown on a wide number of screens.

* **Scatter plot between length and box office performance:**



This scatter plot indicates that there is no obvious linear relationship between a film's length and its box office earnings. For all lengths of videos, there is a spread of data points over the whole y-axis.

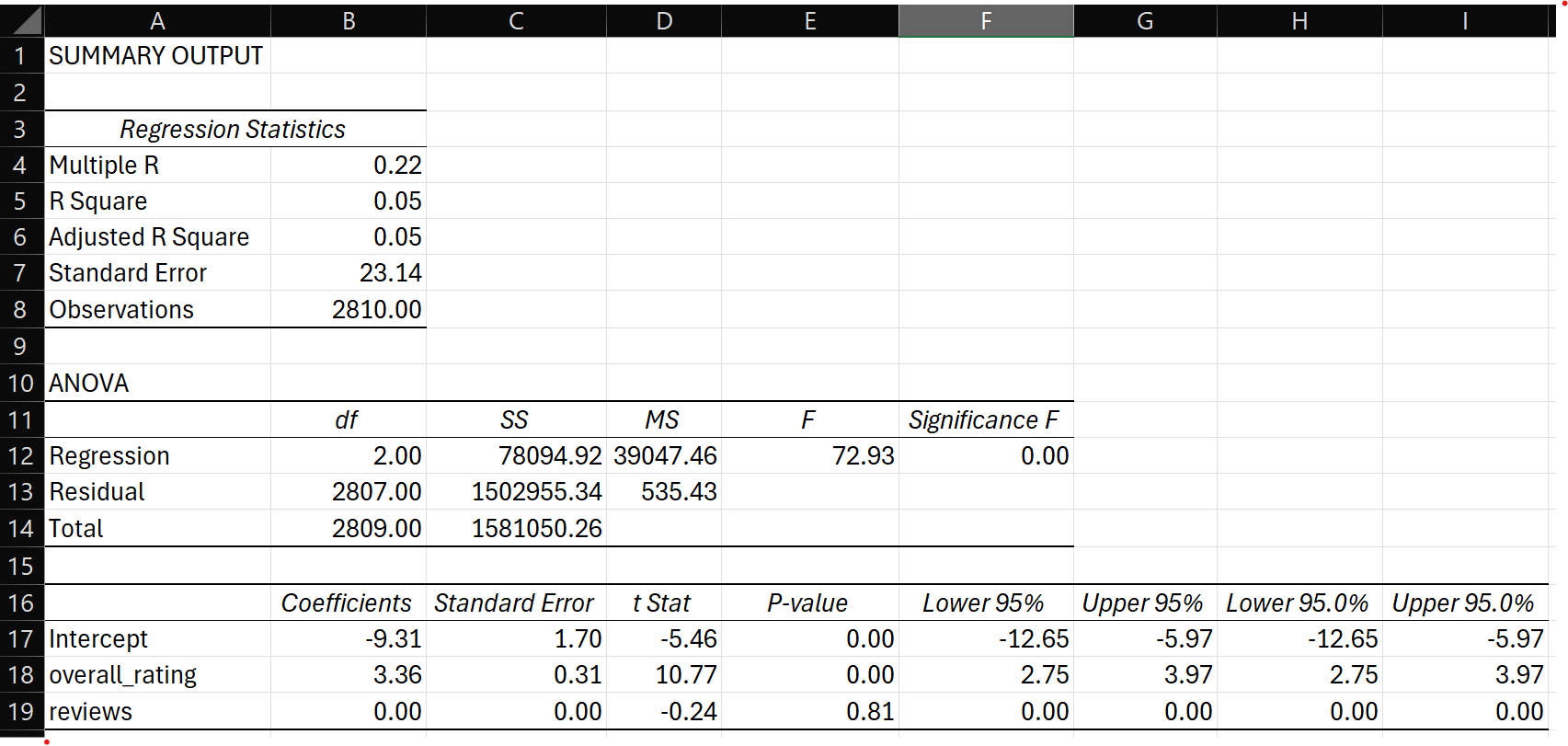
***Correlation between the independent variables and Box office*:**



The correlation table indicates that most independent variables are positively correlated with box office performance. Among them, the number of screens shows the highest correlation (0.810302) with box office performance which we have predicted earlier.

***Regression Analysis*:**

1. **Multiple regression analysis:**
   * We conducted a multiple regression analysis to examine the relationship between overall rating and box office performance, as well as reviews and box office performance. This analysis aimed to determine if there is a significant correlation between these variables.
   * The key factor in identifying a significant relationship is the p-value. If the p-value is less than 0.05, it indicates a statistically significant relationship between the variables. However, if the p-value exceeds 0.05, it suggests that there is no significant relationship among the variables under consideration.



1. ***Overall rating – Box Office performance:***

* From the above sheet diagram, the p value for overall rating is 0.00. It indicates a significant relationship between the overall rating and box office performance. This means there is strong evidence that the observed results are statistically significant and not due to random chance.
* On average, holding all other factors fixed, if the overall rating of a movie is increased by 1 unit, the box office sales are likely to increase by 3360 units. This is based on the coefficient value for the overall rating in the regression analysis.
* The R-Square value, which measures the proportion of variation in the box office performance explained by the variation in the overall rating, is 0.05. This means that about 5% of the variation in box office performance can be explained by the variation in the overall rating for the film. This means the model doesn’t explain data well.
* The standard error (23.14) represents the typical deviation of data points from the regression line's predictions. A larger value implies less accuracy in the model's forecasts, as the observed values tend to be more scattered around the line.

1. ***Reviews – Box Office Performance:***

* The analysis suggests that the number of reviews does not significantly influence a movie's box office performance, as the coefficient for reviews is negligible (0.00) with a high p-value of 0.81, implying that this factor is not a reliable predictor in the given model.
* This suggests that the number of reviews has no significant impact on box office sales, as the coefficient is effectively zero and not statistically significant.
* The R-square value is 0.05.

***Conclusion for this model:***

This study reveals that Overall ratings play a crucial role in determining box office success, with each additional star rating contributing an estimated $3360 to the overall sales. Interestingly, the number of reviews does not seem to have a significant impact on ticket sales. However, it's important to note that the model used in this analysis only accounts for a small portion of the variability in box office performance, indicating other influential factors not considered in this model.

Overall, the regression analysis findings show that the model, while statistically significant, does not explain a significant portion of the variance in 'overall\_ratings' and ‘reviews’. Hence alternative models or new independent variables should be investigated to increase the model's explanatory power.

1. **Multiple Regression Analysis: Tickets and Screens (Predictor Values)**

We performed a regression analysis, including tickets and screens as additional independent variables, revealing the following results. Notably, the p-value for the overall rating is 0.00, indicating a robust association with box office performance.

A screenshot of a graph

Description automatically generated

Based on the coefficient values from the multiple regression analysis:

* On average, holding all other factors fixed, if the pictures/posters is increased by 1 unit, the box office sales are likely to increase by 1000 units.
* On average, holding all other factors fixed, if the length of movie is increased by 1 unit, the box office sales are likely to increase by 140 units.
* On average, holding all other factors fixed, if the number of the trailer is increased by 1 unit, the box office sales are likely to decrease by 320 units.
* On average, holding all other factors fixed, if the ticket prices are increased by 1 unit, the box office sales are likely to increase by 870 units.
* On average, holding all other factors fixed, if the number of screens is increased by 1 unit, the box office sales are likely to increase by 0 units. This might be because screens is a significant predictor of box\_office. So higher screens always results in higher box\_office.

***Tickets:***

* The p-value indicates a significant relationship between ticket with box office performance.
* The R-squared value of 0.67 suggests that approximately 67% of the variation in box office performance can be explained by ticket sales.
* A relatively strong relationship (correlation coefficient of 0.87) exists between ticket and screen numbers, implying their combined influence on box office success.
* The standard error of 13.59 represents the typical deviation of data points from the regression line's predictions, with higher values indicating less accuracy in forecasts due to increased scatter.

***Screens :***

* The standard error of 0.00 suggests that there is no variability in the data points for screens, indicating a perfect fit of the model to the observed data.
* The coefficient of 0.00 indicates that for each unit increase in the number of screens, there is a precisely corresponding increase in box office performance.
* With a p-value of 0.00, screens are confirmed as a highly reliable predictor in the model, exerting a significant influence on box office performance.
* The consistent R-squared value of 0.67 highlights screens as a key factor alongside tickets in explaining box office performance, despite the negligible coefficient value.

1. **Multiple Regression Analysis: Weekend (Predictor Value)**

A screenshot of a computer

Description automatically generated

* The inclusion of the "weekend" variable aims to explore its potential impact on box office performance.
* Movies released during the weekend exhibit a notable effect on box office performance, supported by a p-value of 0.58.
* The coefficient of 0.32 suggests a weak relationship between weekend releases and box office success, indicating a limited influence.
* Despite its inclusion, the R-squared value remains consistent at 0.67, suggesting that nearly 67% of the variation in box office performance can be attributed to other factors.
* The standard error of 13.60 reflects the typical deviation of data points from the regression line's prediction, providing insight into the accuracy of the model's forecasts.
* With a p-value of 0.58, weekend is confirmed as not a reliable predictor in the model, along with award, reviews and overall\_ratings with corresponding p-value 0.71, 0.13 and 0.10 where all these are more than 0.05.

The p-value indicates a significant relationship between pictures, length, trailer, ticket and screens with box office performance.

Based on the coefficient values from the multiple regression analysis:

* On average, holding all other factors fixed, if the pictures/posters is increased by 1 unit, the box office sales are likely to increase by 10 units.
* On average, holding all other factors fixed, if the length of movie is increased by 1 unit, the box office sales are likely to increase by 140 units.
* On average, holding all other factors fixed, if the number of the trailer is increased by 1 unit, the box office sales are likely to decrease by 320 units.
* On average, holding all other factors fixed, if the ticket prices are increased by 1 unit, the box office sales are likely to increase by 860 units.
* On average, holding all other factors fixed, if the number of screens is increased by 1 unit, the box office sales are likely to increase by 0 units. This might be because screens is a significant predictor of box\_office. So higher screens always results in higher box\_office. It never changes between models as it is a strong predictor.

***Conclusion:***

The regression analysis highlights that ticket sales and screen count significantly impact box office performance. These independent variables are anticipated to exert notable effects, positively, owing to their role in driving audience attendance and revenue generation. Furthermore, the introduction of the "weekend" variable implies its potential influence on box office success, albeit modest. In summary, these results offer valuable insights for stakeholders in the film industry, assisting them in strategic decision-making to enhance box office performance.