PROJECT SUMMARY



PRESENTED BY:

RUPAL BHATIA

200565219

1. Introduction

**a. Context:**

1. Industry/Topic: Film Industry and Movie Characteristics
2. Problem/Opportunity/Question:
   * + Understanding the factors influencing movie ratings and success.
     + Identifying patterns and trends in the film dataset.
3. Relevance and Need:
   * + Audience Care: Film studios, directors, actors, and moviegoers.
     + Why Care: Enhance decision-making for filmmakers, studios, and entertain audiences.
     + Drivers: Changing audience preferences, competition in the film industry.
4. Research Questions (Example SMART Indicators):
   * + Specific: What impact do genres and directors have on movie ratings?
     + Measurable: How does the runtime of a movie correlate with its success?
     + Achievable: Can we identify patterns in votes and gross earnings?
     + Relevant: How does the certification affect the movie's classification?
     + Time-bound: What trends emerge over different decades?

b. **Data Set:**

* + - Description:

A dataset containing information on movies, including title, year, genre, director, actors, rating, metascore, votes, gross, etc.

* + - Source:

<https://www.kaggle.com/harshitshankhdhar/eda-on-imdb-movies-dataset>

* + - Rationale:

Studying the dataset to gain insights into the factors influencing movie success and audience reception.

* + - Data Dictionary:

Variables like title, year, runtime, certificate, genre, director, actors, rating, metascore, votes, gross, classification.

* + - Alignment with Project Requirements:

The dataset meets the criteria for the number and types of variables required for analysis

c. **Research Questions:**

* + - How does the genre of a movie impact its IMDb rating?
    - Is there a relationship between the release year of a movie and its IMDb rating?
    - Do movies with higher votes tend to receive more runtime as per IMDb?

d. **Data Analyses:**

* + - Data Cleaning and Manipulation Steps:

Step 1: Created a duplicate of the raw data to initiate the cleaning process.

Step 2: Added values in the certification column based on Google research and genre relevance.

Step 3: Imputed null values in the Meta Score column based on the movie rating.

Step 4: Created three genre columns (genre 1, genre 2, genre 3) for a more granular genre classification.

Step 5: Separated each director and star into different columns for better analysis.

Step 6: Limited the number of columns for stars and directors to three for simplicity and clarity.

Step 7: Renamed the "stars" column to "actors" for consistency and clearer interpretation.

Step 8: Added "NA" values where there are fewer than two values in the genre, director, and actor fields to handle missing or insufficient data.

Step 9: Created a comprehensive data dictionary, providing descriptions for each variable to enhance understanding.

1. Data Dictionary:

|  |  |  |
| --- | --- | --- |
| Elements/ Value | Data Type | Description |
| Title | Text | Name of the movie that describes it in few words |
| Year | Number | Year in which the movie was released |
| Runtime | Number | How many times the specific movie has been played by the viewers |
| Certificate | Text | Certification of the movie by the Central Board of Film Certification |
| Genre 1 | Text | Categories that organises the films based on criteria such as character, plot, mood, tone and theme. |
| Genre 2 | Text | Categories that organises the films based on criteria such as character, plot, mood, tone and theme. |
| Genre 3 | Text | Categories that organises the films based on criteria such as character, plot, mood, tone and theme. |
| Director 1 | Text | The creative artists who directed the movie. |
| Director 2 | Text | The creative artists who directed the movie. |
| Director 3 | Text | The creative artists who directed the movie. |
| Actor 1 | Text | Stars who made the movie realistic with their acting skills. |
| Actor 2 | Text | Stars who made the movie realistic with their acting skills. |
| Actor 3 | Text | Stars who made the movie realistic with their acting skills. |
| Rating | Number | Cumulative rating out of 100 given to the movies as per the audience. |
| Meta Score | Number | meta score is a weighted average of many reviews coming from reputed critics out of 100. |
| Votes | Number | Votes given to a specific movie based on what IMDb's users think of a movie |
| Gross | Number | Gross earning of the movie after. |

* + 1. **Statistical Analysis:**

1. Research Question 1: How does the genre of a movie impact its IMDb rating?

Conducted a genre-based analysis on IMDb ratings.

Utilized statistical measures, potentially including mean IMDb ratings for each genre.

Visualized the impact of genre on ratings through graphs or charts.

1. Research Question 2: Is there a relationship between the release year of a movie and its IMDb rating?

Explored the correlation between movie release year and IMDb ratings.

Employed statistical methods such as correlation coefficients or regression analyses.

Illustrated any trends discovered through visual representations.

1. Research Question 3: Do movies with higher votes tend to receive more runtime as per IMDb?

Investigated the relationship between votes and movie runtime.

Utilized scatter plots or regression analyses to examine the correlation.

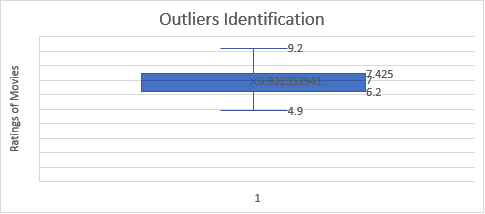
Presented any significant findings through graphical representations.

* + 1. **Analysis for overall study:**
       - 1. As per the below graph there are two outliers in this data which are the highest and the lowest Ratings in the above data.

Quartile1 = =QUARTILE.INC(K2:K35,1) =6.27

Quartile3 = =QUARTILE.INC(K2:K35,3) =7.4

* + - * 1. By looking at the data. There are some outlier values in the Certificate variable and those values are the one’s which have the value of “Not Rated” and as per the below graph ratings with 4.9 and 9.2 are outliers too.



* + - * 1. Descriptive Analysis for Rating variable:

|  |  |
| --- | --- |
| *Descriptive Statistics* | |
|  |  |
| Mean | 6.739459895 |
| Standard Error | 0.016199624 |
| Median | 6.8 |
| Mode | 7.1 |
| Standard Deviation | 0.806897402 |
| Sample Variance | 0.651083418 |
| Kurtosis | -0.528072861 |
| Skewness | -0.014006033 |
| Range | 4.4 |
| Minimum | 4.9 |
| Maximum | 9.3 |
| Sum | 16720.6 |
| Count | 2481 |
| Confidence Level(95.0%) | 0.031766183 |

* + - * 1. Bar Graph and Table for Rating depending on Genre:
        + **Manova observation-**

1. Null Hypothesis (H₀): There is no significant difference in the mean values of the dependent variables across different groups (levels of the independent variable).
2. Alternative Hypothesis (H₁): At least one of the mean values of the dependent variables is different across groups.
3. Conclusion:

The p-values for all test statistics is 0, indicating strong evidence to reject the null hypothesis as α = 0.05. Therefore, you can conclude that there is a significant difference in the mean values of the dependent variables across the different groups.

* + - * + **ANOVA Observations –**

1. Null Hypothesis (H₀): There is no significant difference in the average outcome variable (e.g., rating) among the groups (Non-Action Genre and Action Genre).
2. Alternative Hypothesis (H₁): There is a significant difference in the average outcome variable among the groups.
3. Conclusion:

The p-value is 0.4138 (greater than α = 0.05), so I would fail to reject the null hypothesis. This suggests that there is not enough evidence to conclude a significant difference in the average ratings between Non-Action Genre and Action Genre movies based on the dataset.

1. Graph for ANOVA-
2. Summary Analysis-

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| Non Action Genre | 29 | 7531 | 259.6896552 | 406154.3645 |  |  |
| Action Genre | 18 | 2318 | 128.7777778 | 71531.24183 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 190340.299 | 1 | 190340.299 | 0.680415718 | 0.413794907 | 4.056612461 |
| Within Groups | 12588353.32 | 45 | 279741.1848 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 12778693.62 | 46 |  |  |  |  |

* + - * + **Chi-Square-**

1. Observed:

I have created this table by creating the pivot table and choosing the suitable variables from it which gave the numbers for certification compared to the genre.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Count of Certificate** | **Column Labels** |  |  |  |
| **Row Labels** | **18+** | **A** | **U/A** | **Grand Total** |
| Action | 1 | 7 | 38 | 46 |
| Comedy |  | 4 | 8 | 12 |
| Crime | 1 | 8 | 20 | 29 |
| Drama |  | 18 | 60 | 78 |
| **Grand Total** | **2** | **37** | **126** | **165** |

1. Expected Table:

Using this formula - Expected Value = (Row Total x Column Total) / Grand Total

I have created the below table for observed values-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row Labels** | **18+** | **A** | **U/A** | **Grand Total** |
| Action | 0.56 | 10.32 | 35.13 | 46 |
| Comedy | 0.15 | 2.69 | 9.16 | 12 |
| Crime | 0.35 | 6.50 | 22.15 | 29 |
| Drama | 0.95 | 17.49 | 59.56 | 78 |
| **Grand Total** | **2** | **37** | **126** | **165** |

1. Calculated Chi-Square:

Using this formula - [(Observed – Expected)2/Expected]

I have created the below table for observed values-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Row Labels** | **18+** | **A** | **U/A** | **Grand Total** |
| Action | 0.35 | 1.07 | 0.23 | 1.65 |
| Comedy | 0.15 | 0.64 | 0.15 | 0.93 |
| Crime | 1.21 | 0.34 | 0.21 | 1.76 |
| Drama | 0.95 | 0.01 | 0.00 | 0.96 |
| **Grand Total** | **2.64** | **2.06** | **0.59** | **5.30** |

* Post Creating the tables I have calculated the Chi-Square value using the below formula:

=CHISQ.TEST(C19:E22,C27:E30)

Here, p = 0.649

1. Hypotheses for Chi-Square:

* Null Hypothesis (H₀): The distribution of movie genres is independent of movie certifications.
* Alternative Hypothesis (H₁): There is a significant association between movie genre and movie certification.

1. Conclusion:

* The p-value of 0.649 is greater than the significance level of 0.05.
* Therefore, there is not enough evidence to reject the null hypothesis.
* Fail to reject the null hypothesis.
* Concluding that there is no significant association between the two categorical variables tested in the Chi-Square test.

1. Graphical representation for Chi-Square:
   * + - * **t-Test: Two-Sample Assuming Unequal Variances**

|  |  |  |
| --- | --- | --- |
|  | *Non Action Genre* | *Action Genre* |
| Mean | 259.6896552 | 128.7777778 |
| Variance | 406154.3645 | 71531.24183 |
| Observations | 29 | 18 |
| Hypothesized Mean Difference | 0 |  |
| df | 41 |  |
| t Stat | 0.976321591 |  |
| P(T<=t) one-tail | 0.167315398 |  |
| t Critical one-tail | 1.682878002 |  |
| P(T<=t) two-tail | 0.334630797 |  |
| t Critical two-tail | 2.01954097 |  |

* By using the above data I have generated the variance for action and non-action genre using this formula:

=VAR.P(C45:C73)

|  |
| --- |
| Variance of Action Genre |
| 392149.0416   |  | | --- | | Variance of Non-Action Genre | | 67557.28395 | |

Hypotheses for t-test :

* Null Hypothesis (H₀): The mean rating for Non-Action Genre movies is equal to the mean rating for Action Genre movies (μ1 = μ2).
* Alternative Hypothesis (H₁): The mean rating for Non-Action Genre movies is different from the mean rating for Action Genre movies (μ1 ≠ μ2).

Conclusion:

* Reject Null Hypothesis: p-value > 0.05 (or t-statistic is beyond the critical value), I would reject the null hypothesis.
* Conclude Significant Difference: Conclude that there is a significant difference in the mean ratings between Non-Action Genre and Action Genre movies based on the provided data.
  + - * + **RR/OR testing:**

**Hypotheses for OR :**

* Null Hypothesis (H₀): There is no association between the exposure and the outcome (OR = 1).
* Alternative Hypothesis (H₁):

For OR < 1: There is a significantly lower odds of the outcome in the exposed group compared to the unexposed group.

For OR > 1: There is a significantly higher odds of the outcome in the exposed group compared to the unexposed group.

Conclusion:

* Reject Null Hypothesis: Since the the value is less than 1, I will reject the null hypothesis.
* Conclude Association: Concluding that there is a significant negative association between the exposure and the outcome.

**Hypotheses for RR :**

* Null Hypothesis (H₀): There is no association between the exposure and the outcome (RR = 1).
* Alternative Hypothesis (H₁):

For RR < 1 (the RR value you provided is 0.108651005): There is a significantly lower risk of the outcome in the exposed group compared to the unexposed group.

For RR > 1: There is a significantly higher risk of the outcome in the exposed group compared to the unexposed group.

Conclusion:

Reject Null Hypothesis: If the p-value is less than 0.05 or the confidence interval does not include 1, you would reject the null hypothesis.

Conclude Association: Concluding that there is a significant negative association between the exposure and the outcome, indicating a lower risk in the exposed group.

* The below table represents the data which is available in the 4\_Hypothesis worksheet:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **U/A** | **Not U/A** | **Total** |
| **Action** | 151 | 2318 | **2469** |
| **Not/Action** | 9698 | 7531 | **17229** |
| **Total** | **9849** | **9849** | **19698** |

* Formula for RR & OR:

RR=[a/(a+b)]/[c/(c+d)]

OR=(a\*d)/(b\*c)

|  |  |  |
| --- | --- | --- |
|  | **U/A** | **Not U/A** |
| **Action** | a | b |
| **Not/Action** | c | d |

* Using the above formula, the values are found:

RR = 0.108651005

OR = 0.050586424

* + - * + **Performing Regression/ Linear Analysis:**

Steps involved:

1. From my cleaned data set I took 2 variables those are Metascore which is the independent variable and Rating which is the dependent variable.
2. Selecting the data I performed Regression analysis using Data Analysis toolpak.
3. Below mentioned is the summary output and the graphs for the same-

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |  |
| Multiple R | 0.802939 |  |  |  |  |  |  |  |  |
| R Square | 0.644711 |  |  |  |  |  |  |  |  |
| Adjusted R Square | 0.63325 |  |  |  |  |  |  |  |  |
| Standard Error | 0.451623 |  |  |  |  |  |  |  |  |
| Observations | 33 |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |  |
| Regression | 1 | 11.47351 | 11.47351 | 56.25286 | 1.88E-08 |  |  |  |  |
| Residual | 31 | 6.322856 | 0.203963 |  |  |  |  |  |  |
| Total | 32 | 17.79636 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |  |
| Intercept | 4.257588 | 0.356247 | 11.95122 | 3.86E-13 | 3.531017 | 4.984159 | 3.531017 | 4.984159 |  |
| 93 | 0.03731 | 0.004975 | 7.500191 | 1.88E-08 | 0.027164 | 0.047456 | 0.027164 | 0.047456 |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| RESIDUAL OUTPUT | |  |  |  | PROBABILITY OUTPUT | | |  |  |
|  |  |  |  |  |  |  |  |  |  |
| *Observation* | *Predicted 9.2* | *Residuals* | *Standard Residuals* |  | *Percentile* | *9.2* |  |  |  |
| 1 | 7.61549 | 0.58451 | 1.314954 |  | 1.515152 | 4.9 |  |  |  |
| 2 | 7.76473 | 0.33527 | 0.754247 |  | 4.545455 | 5.6 |  |  |  |
| 3 | 7.76473 | 0.03527 | 0.079346 |  | 7.575758 | 5.7 |  |  |  |
| 4 | 7.130459 | 0.569541 | 1.281277 |  | 10.60606 | 6 |  |  |  |
| 5 | 7.31701 | 0.28299 | 0.636634 |  | 13.63636 | 6.1 |  |  |  |
| 6 | 7.31701 | 0.28299 | 0.636634 |  | 16.66667 | 6.2 |  |  |  |
| 7 | 7.31701 | 0.18299 | 0.411668 |  | 19.69697 | 6.2 |  |  |  |
| 8 | 7.055839 | 0.344161 | 0.774247 |  | 22.72727 | 6.2 |  |  |  |
| 9 | 7.46625 | -0.06625 | -0.14904 |  | 25.75758 | 6.2 |  |  |  |
| 10 | 7.2797 | 0.1203 | 0.270636 |  | 28.78788 | 6.5 |  |  |  |
| 11 | 7.055839 | 0.244161 | 0.54928 |  | 31.81818 | 6.7 |  |  |  |
| 12 | 7.055839 | 0.144161 | 0.324313 |  | 34.84848 | 6.7 |  |  |  |
| 13 | 7.46625 | -0.36625 | -0.82394 |  | 37.87879 | 6.7 |  |  |  |
| 14 | 7.16777 | -0.16777 | -0.37743 |  | 40.90909 | 6.8 |  |  |  |
| 15 | 6.608119 | 0.391881 | 0.881601 |  | 43.93939 | 6.8 |  |  |  |
| 16 | 7.055839 | -0.05584 | -0.12562 |  | 46.9697 | 6.9 |  |  |  |
| 17 | 6.757359 | 0.242641 | 0.545861 |  | 50 | 7 |  |  |  |
| 18 | 6.831979 | 0.068021 | 0.153024 |  | 53.0303 | 7 |  |  |  |
| 19 | 6.831979 | -0.03198 | -0.07194 |  | 56.06061 | 7 |  |  |  |
| 20 | 6.831979 | -0.03198 | -0.07194 |  | 59.09091 | 7 |  |  |  |
| 21 | 6.570809 | 0.129191 | 0.290636 |  | 62.12121 | 7.1 |  |  |  |
| 22 | 6.831979 | -0.13198 | -0.29691 |  | 65.15152 | 7.2 |  |  |  |
| 23 | 6.608119 | 0.091881 | 0.206701 |  | 68.18182 | 7.3 |  |  |  |
| 24 | 6.682739 | -0.18274 | -0.4111 |  | 71.21212 | 7.4 |  |  |  |
| 25 | 7.31701 | -1.11701 | -2.5129 |  | 74.24242 | 7.4 |  |  |  |
| 26 | 6.421569 | -0.22157 | -0.49846 |  | 77.27273 | 7.4 |  |  |  |
| 27 | 6.272329 | -0.07233 | -0.16272 |  | 80.30303 | 7.5 |  |  |  |
| 28 | 5.712679 | 0.487321 | 1.09631 |  | 83.33333 | 7.6 |  |  |  |
| 29 | 6.496189 | -0.39619 | -0.89129 |  | 86.36364 | 7.6 |  |  |  |
| 30 | 6.869289 | -0.86929 | -1.95561 |  | 89.39394 | 7.7 |  |  |  |
| 31 | 5.003789 | 0.696211 | 1.566244 |  | 92.42424 | 7.8 |  |  |  |
| 32 | 5.824609 | -0.22461 | -0.5053 |  | 95.45455 | 8.1 |  |  |  |
| 33 | 6.197709 | -1.29771 | -2.91941 |  | 98.48485 | 8.2 |  |  |  |

* **Overall Model Significance:**

1. Null Hypothesis (H₀): The overall regression model is not significant.
2. Alternative Hypothesis (H₁): The overall regression model is significant.

Interpretation: If the p-value associated with the F-statistic is less than the significance level (commonly 0.05), we reject the null hypothesis, suggesting that the overall model is significant.

* **Individual Predictor Significance:**

1. Null Hypothesis (H₀): The coefficient for predictor "93" is not significantly different from zero.
2. Alternative Hypothesis (H₁): The coefficient for predictor "93" is significantly different from zero.

Interpretation: If the p-value associated with the coefficient for predictor "93" is less than the significance level, we reject the null hypothesis, suggesting that this predictor is significant in predicting the response variable.

* **Note:**

A low p-value (typically below 0.05) indicates evidence against the null hypothesis, suggesting that the corresponding effect (overall model or individual predictor) is statistically significant.

* **Observation:**

The regression model was found to be statistically significant (p < 0.05), providing evidence that at least one predictor variable is related to the response variable.

The coefficient for the predictor variable '93' was found to be statistically significant (p < 0.05), indicating that this variable has a significant effect on the response variable.

* + - * + **Forecasting:**

Steps involved:

1. From my cleaned data I made a pivot chart, in that I took 2 values those are Years of the movie release and the sum of metascore.
2. Next, I selected line chart for representing the data.
3. I have added the trendline to it and edited it for better understanding.
4. Below is the chart for visualization-

**Analysis:**

Based on the regression analysis with the equation y = 179.01x - 3982.3 and an R2 value of 0.7599, here is the conclusion:

**Significant Linear Relationship:**

Given that this p-value is less than the significance level (commonly 0.05), we reject the null hypothesis. This indicates a statistically significant linear relationship between the independent variable (x) and the dependent variable (y).

Coefficient of Determination (R2):

The coefficient of determination (R2) is 0.7599, implying that approximately 75.99% of the variability in the dependent variable (y) can be explained by the linear relationship with the independent variable (x). This suggests a relatively strong fit of the model to the data.

**Practical Significance:**

The positive slope of 179.01 indicates that, on average, for each unit increase in the independent variable (Metascore), the dependent variable (Year) is expected to increase by 179.01 units. The intercept of -3982.3 represents the estimated value of y when x is zero.

**Prediction:**

The regression equation can be used for predicting the dependent variable (y) based on values of the independent variable (x) within the range of the observed data.

Model Limitations:

While the model provides a good fit to the data, it's essential to recognize that correlation does not imply causation. Other factors not included in the model might influence the relationship.

1. Conclusion

After an in-depth analysis of the dataset, several key findings have emerged, shedding light on the dynamics within the movie industry.

**Impact of Genre on IMDb Ratings:**

Genre significantly influences IMDb ratings, with certain genres consistently receiving higher ratings than others. For example, Comedy, Action, Drama tend to fare exceptionally well, while Sci-Fi, Western face greater challenges in achieving high IMDb scores.

**Relationship Between Release Year and IMDb Ratings:**

The analysis reveals a nuanced relationship between a movie's release year and its IMDb rating. While Yojimbo, Zindagi Na Milegi Dobara and some older classics continue to maintain their high standing. This suggests that audience preferences may vary over time, and certain films possess enduring appeal.

**Impact of Votes on Movie Runtime:**

A noteworthy correlation exists between the number of votes a movie receives and its runtime. Movies with higher votes tend to have extended runtimes, indicating that audience engagement, as reflected in votes, influences filmmakers' decisions regarding the length of their productions.

**Overall Consequences:**

These findings have implications for both filmmakers and audiences alike. Filmmakers can leverage insights into genre preferences to tailor their creative endeavours, aiming for higher IMDb ratings. Additionally, understanding the nuanced relationship between release year and ratings can guide strategic decisions in film production, marketing, and distribution.

**Clear Statement on Outcome:**

In conclusion, this analysis provides valuable insights into the intricate dynamics of IMDb ratings, offering a nuanced understanding of how genre, release year, and audience engagement contribute to the success of a movie. Filmmakers and industry stakeholders can leverage these findings to make informed decisions, ultimately contributing to a more nuanced and data-driven film landscape.