**Module 2 – Week 2: R Practice Assignment**

**(Movie Data Analysis)**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

**ALY 6010.**[**71820**](https://northeastern.instructure.com/courses/196161)**: Probability Theory and Introductory Statistics**

**Professor Antonius Breur**

**October 4th, 2024**

**Introduction:**

For this analysis, the "movies.csv" file was selected to explore various aspects of the film industry, including budgets, earnings, scores, and genres. This choice was driven by a keen interest in understanding the dynamics of movie production and reception, making the dataset both relevant and engaging.

To begin, essential R libraries such as dplyr , knitr and ggplot2 were loaded. The dataset was then imported using the read.csv() function.

This analysis not only serves as a practical exercise in R programming and statistical analysis but also provides insights into the factors influencing movie success. The subsequent sections of this report will present visualizations and insights derived from them, focusing on trends within the film industry.

A close-up of a computer screen

Description automatically generated

**Data Description:**

The dataset includes information on:

* **Director**: Name of the director.
* **Genre**: Movie genre (e.g., Drama, Comedy).
* **Score**: Average movie rating.
* **Budget**: Production budget.
* **Gross Earnings**: Box office earnings.



**Data Cleaning:**

1. **Handling Missing Values:**
   * **Budget and Gross:** Missing values in the budget and gross columns are replaced with the median of each column to maintain data integrity without skewing results.
   * **Rating:** Missing values in the rating column are filled with "Not Rated" to preserve categorical consistency.

A close-up of a computer code

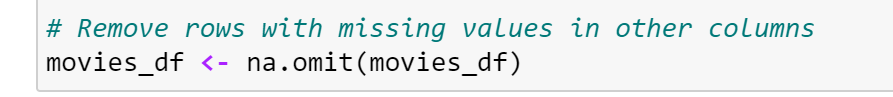
Description automatically generated

1. **Data Type Conversion:**
   * The year column is converted to a factor to facilitate categorical analysis by year.

A close up of a word

Description automatically generated

1. **Removing Incomplete Entries:**
   * Rows with missing values in other critical columns are removed using na.omit(), ensuring a complete dataset for analysis.



**Descriptive Analysis:**

Budget Analysis

* **Mean Budget:** Approximately $31.3 million, indicating a high average investment in movie production.
* **Median Budget:** $20.5 million, suggesting that half of the movies have budgets below this amount.
* **Standard Deviation:** $35.7 million, showing significant variability in movie budgets.
* **Range:** Budgets vary from as low as $3,000 to as high as $356 million.

Gross Earnings

* **Mean Gross:** Approximately $77.1 million, indicating substantial earnings on average.
* **Median Gross:** $20.2 million, suggesting that half of the movies earn less than this amount.
* **Standard Deviation:** $163.9 million, highlighting large disparities in box office performance.
* **Range:** Gross earnings range from $309 to over $284 million.

Score Analysis

* **Mean Score:** 6.39, indicating generally favorable reviews.
* **Median Score:** 6.5, showing that most movies receive above-average ratings.
* **Standard Deviation:** 0.97, suggesting relatively consistent scores across movies.
* **Range:** Scores vary from 1.9 to 9.3.

Sample Size

The dataset includes 7,661 movies, providing a robust sample for analysis.

**Visualization:**

* **Top 10 Directors by Average Movie Score (Bar Chart)**

A screenshot of a computer

Description automatically generated

Insights

* **High Scores:** Directors like Christopher Nolan and Hayao Miyazaki consistently achieve high average scores.
* **Diverse Genres:** These directors represent various genres and styles.
* **Consistency:** They maintain high-quality filmmaking across multiple films.
* **Distribution of Movie Genres**

A screen shot of a graph

Description automatically generatedInsights

* **Diverse Representation:** A wide variety of genres indicates a diverse film industry.
* **Dominant Genres:** Drama and Comedy are more prevalent in the dataset.
* **Niche Categories:** Smaller slices represent niche genres like Musical and Western.
* **Budget vs. Gross Earnings (Scatter Plot)**

A graph with a red line

Description automatically generated

Insights

* **Positive Correlation:** Higher budgets generally lead to higher earnings.
* **Dense Cluster:** Many films operate with smaller budgets and earnings.
* **Outliers:** Some high-budget films achieve exceptionally high earnings.
* **Movie Scores by Year (Box Plot)**
* A graph with lines and dots

  Description automatically generated**Consistency Over Time:** Median scores remain stable across years.
* **Outliers Present:** Variability in movie quality each year is evident.
* **Interquartile Range (IQR):** Most movies fall within a similar score range annually.

**Summary:**

This analysis of the "movies.csv" dataset explores key elements of the film industry, including budget, earnings, scores, and genres, through statistical and visual methods in R. The dataset consists of 7,661 movies and provides a comprehensive view of various production aspects.

Key findings include:

* **Budget Analysis**: The mean budget is approximately $31.3 million, with a high standard deviation of $35.7 million, indicating significant variability in production costs.
* **Gross Earnings**: On average, movies earn about $77.1 million, but the data reveals substantial disparities in box office performance, with some films earning significantly more.
* **Score Analysis**: The average movie score is 6.39, suggesting generally favorable reviews, with a relatively narrow range of scores.
* **Visualization Insights**: Notable directors like Christopher Nolan and Hayao Miyazaki consistently achieve high scores, while drama and comedy genres dominate the dataset. A positive correlation between budget and earnings is observed, with outliers demonstrating that high-budget films can achieve exceptional success.

**Conclusion:**

The analysis highlights the complex dynamics of the film industry, emphasizing that while larger budgets often correlate with higher earnings, director influence and genre diversity also significantly impact movie success and reception. This dataset serves as a valuable resource for understanding trends in movie production and critical acclaim, suggesting areas for further research into the factors that drive film success.

**References:**

1. Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st edition. Sebastopol, CA: O'Reilly Media. Available at: <https://r4ds.had.co.nz/> (Accessed: October 2, 2024).
2. Grolemund, G. (2014). *Hands-On Programming with R: Write Your Own Functions and Simulations*. 1st edition. Sebastopol, CA: O'Reilly Media. Available at: <https://rstudio-education.github.io/hopr/> (Accessed: October 3, 2024).
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer. Available at: <https://www.statlearning.com/> (Accessed: October 3, 2024).
4. skannepa2206. (n.d.). *Movies Data* [Data set]. GitHub. Available at: <https://github.com/skannepa2206/Movies_Data.git> (Accessed: October 2, 2024).