The International Journal of Embedded Computer Engineering (IJEmCE), Vol. 1, No. 1, April 2026: 1–51
https://ejurnal.ung.ac.id/index.php/IJEmCE

Exploratory Data Analysis on TMDb Top Rated Movies Dataset Using Winsorization Approach

M. Fauzan Syaifullah[1*], Ulfatun Nadifa[2], Wahab Musa[3], Rahmat Hidayat Dongka[4], Ade Irawaty Tolango[5], Zainudin Bonok[6], Ikhsan Hidayat[7]
Computer Engineering, Universitas Negeri Gorontalo, Indonesia[1,2,3,7]
Electrical Engineering, Universitas Negeri Gorontalo, Indonesia[4,5,6]
* 268fauzansyaifullah@gmail.com

Abstract -- This research presents an Exploratory Data Analysis (EDA) on the top-rated movies dataset from The Movie Database (TMDb) spanning 1902–2026. The main objective of this study is to clean the data, identify distribution biases, and prepare the dataset for predictive modeling. The approach includes missing value imputation, skewness metric evaluation, and correlation analysis. Findings reveal highly positive skewness in popularity and vote count variables, as well as a temporal bias dominated by modern era movies. To handle extreme values (outliers), the Interquartile Range (IQR) method combined with the capping (Winsorization) technique was applied. As a result, the data distribution became more stable without losing representative information from blockbuster movies. Correlation analysis revealed a strong positive relationship (0.62) between popularity and vote count, and multicollinearity (0.97) between the month and quarter variables, which needs to be eliminated in the subsequent machine learning phase.

Keywords:
EDA;
IQR;
Dataset TMDb;
Movie;
Winsorization;

Article History:
Received: March 15, 2026
Revised: April 22, 2026
Accepted: April 28, 2026
Published: April 30, 2026

DOI: 10.xxxxx/ijemce.v1i1.xxxxx

I. INTRODUCTION

The global film industry generates massive volumes of data each year, including popularity metrics, audience ratings, and commercial performance indicators. Understanding the hidden patterns within this data is essential for developing recommendation systems and predictive models for movie success. However, in the application of machine learning algorithms, the quality of raw data remains a major challenge, as it often contains missing values and extreme values (outliers) [1]. The presence of outliers, particularly in datasets with highly right-skewed distributions, can significantly distort statistical analysis and degrade the interpretability of predictive models [2]. In the Top Rated Movies from TMDb (1902–2026) dataset, the primary issue identified is distributional bias in popularity metrics and vote counts. If these outliers are ignored, or conversely removed entirely through trimming, the model may lose critical information that represents real-world phenomena, such as blockbuster films that naturally generate extreme values.

Therefore, a more robust data preprocessing approach is required. Previous studies have demonstrated that the Winsorization technique (or capping) is effective in improving the performance and stability of machine learning algorithms when dealing with outlier-prone data, without discarding important observations [3]. This approach limits extreme values to specific percentile thresholds, resulting in a more stable and representative data distribution [4]. In addition, Exploratory Data Analysis (EDA) plays a crucial role in understanding the structure, distribution, and quality of datasets before applying advanced analytical models. According to John W. Tukey, EDA enables researchers to uncover patterns, detect anomalies, and test assumptions through visualization and summary statistics [6]. Effective EDA helps ensure that preprocessing steps are grounded in empirical observations rather than arbitrary decisions.

Moreover, handling skewed data distributions is particularly important in real-world datasets such as movie ratings and popularity scores, where a small number of observations dominate the overall distribution. Techniques such as log transformation, normalization, and robust scaling are often applied; however, Winsorization provides a balanced approach by preserving the dataset size while mitigating

extreme influence [7]. Another critical aspect in data preprocessing is addressing multicollinearity, which occurs when independent variables are highly correlated. Multicollinearity can reduce model stability and lead to unreliable coefficient estimates, particularly in regression-based models [8]. Correlation analysis and dimensionality reduction techniques are therefore essential steps in preparing high-quality datasets.

Furthermore, recent advances in data science emphasize the importance of data-centric approaches, where improving data quality can lead to better model performance than merely optimizing algorithms [9]. This perspective reinforces the significance of robust preprocessing techniques such as Winsorization and systematic EDA in achieving reliable and generalizable results. Finally, the increasing availability of large-scale movie datasets from platforms such as TMDb highlights the need for scalable and efficient analytical frameworks. These datasets provide valuable insights into audience behavior, content trends, and market dynamics, making them ideal for experimentation in machine learning and data analysis [10][16].

This study focuses on conducting a comprehensive Exploratory Data Analysis (EDA) on the TMDb dataset. The main objectives of this research are: (1) to identify data distribution characteristics and temporal bias, (2) to handle outliers effectively using the Interquartile Range (IQR) method combined with Winsorization, and (3) to analyze correlations among variables to mitigate the risk of multicollinearity. Through these steps, the dataset is ensured to be clean, balanced, and ready for further analytical modeling stages [5].

II. METHOD

This study adopts a quantitative approach using Exploratory Data Analysis (EDA) to identify patterns, anomalies, and relationships among variables in the Top Rated Movies from TMDb dataset. The dataset spans a long temporal range from 1902 to 2026, enabling both cross-sectional and temporal analysis. The overall workflow consists of several structured stages to ensure data quality, robustness, and analytical reliability.

1. Data Collection and Understanding

   The dataset includes key attributes such as movie popularity, vote count, ratings, release year, and other metadata. An initial data inspection is conducted to understand the structure, data types, and distribution of each variable. Descriptive statistics (mean, median, standard deviation) are used to provide a preliminary overview of the dataset [11].

2. Data Cleaning

   Data cleaning is a crucial step to ensure the dataset is suitable for analysis. The following techniques are applied:

   a. Handling Missing Values:
      1. Numerical variables are imputed using the median, as it is more robust to extreme values compared to the mean. This ensures that the central tendency is not biased by outliers.
      2. Categorical variables with missing values are filled with an empty string or a placeholder label to preserve dataset consistency.
   b. Data Type Adjustment: Variables are converted into appropriate data types (e.g., date, numeric, categorical) to support accurate computation and analysis.
   c. Duplicate Removal: Duplicate records, if any, are identified and removed to prevent bias in statistical analysis.

3. Univariate Analysis

   Univariate analysis is conducted to examine the distribution of individual variables. This includes:

   1. Histogram and density plots to observe distribution shapes.
   2. Identification of skewness (especially right-skewed distributions common in popularity and vote count).
   3. Summary statistics to detect unusual patterns This step provides insight into the nature of the data before applying more advanced techniques.

4. Outlier Detection

   Outliers are identified using both visual and statistical approaches:

   a. Boxplot Visualization: Used to visually detect extreme values and distribution spread.
   b. Interquartile Range (IQR) Method: Outliers are determined using the following formula:
      Lower Bound = $Q1 - 1.5 \times IQR$
      Upper Bound = $Q3 + 1.5 \times IQR$
      where $IQR = Q3 - Q1$. Values outside these bounds are considered outliers [11].

   This dual approach ensures accurate and reliable detection of extreme observations.

5. Outlier Treatment (Winsorization Approach)

   Instead of removing outliers (trimming), this study applies Winsorization (capping):

a. Values exceeding the upper bound are replaced with the upper threshold.
b. Values below the lower bound are replaced with the lower threshold.

This approach is chosen because:

a. It preserves important observations (e.g., blockbuster movies with extreme popularity).
b. It reduces the impact of extreme values on statistical analysis.
c. It improves model stability without reducing dataset size.

Winsorization ensures a more balanced and representative distribution while maintaining real-world relevance [12].

6. Bivariate and Multivariate Analysis
   a. Correlation Analysis
      To examine relationships between variables, Pearson correlation coefficient is used. The results are visualized using a heatmap, allowing easy identification of strong positive or negative relationships.
   b. Multicollinearity Detection
      Highly correlated variables are identified to prevent redundancy and instability in predictive modeling. If necessary, feature selection or dimensionality reduction techniques can be applied.
   c. Feature Interaction Exploration
      Scatter plots and pair plots are used to analyze interactions between key variables such as popularity, vote count, and rating [13].
7. Temporal Analysis
   Given the wide time span of the dataset, temporal analysis is performed to identify trends over time, such as:
   a. Changes in movie popularity across decades.
   b. Evolution of audience ratings.
   c. Growth patterns in vote counts.

   This step helps uncover long-term patterns and industry shifts.
8. Data Validation and Final Preparation
   After all preprocessing steps, the dataset is re-evaluated to ensure:
   a. No missing values remain
   b. Distribution is stable and balanced
   c. No extreme outliers distort the analysis
   d. Features are ready for modeling
9. Tools and Technologies
   The analysis is implemented using:
   a. Python programming language
   b. Libraries such as Pandas, NumPy, Matplotlib, and Seaborn
   c. Statistical techniques integrated within the data science workflow

Overall, this methodology emphasizes a robust and systematic data preprocessing pipeline, combining statistical techniques (IQR, correlation analysis) with practical approaches (Winsorization) to ensure high-quality data. This structured approach enables more reliable insights and prepares the dataset for future machine learning modeling [14].

III. RESULT AND DISCUSSION

A. Distribution Analysis and Temporal Bias

The initial distribution analysis reveals that key numerical variables, particularly popularity and vote_count, exhibit a highly right-skewed distribution. This indicates that the majority of movies are concentrated at lower values, while only a small proportion achieves exceptionally high popularity or vote counts. Such a distribution is typical in real-world datasets influenced by the “long-tail phenomenon”, where a few items (e.g., blockbuster films) dominate attention and engagement, while the majority remain relatively obscure [15].

From a statistical perspective, this extreme skewness has important implications. Measures of central tendency such as the mean become less representative of the dataset, as they are heavily influenced by extreme values. In contrast, the median provides a more robust estimate of the central location. The presence of this skewness also suggests potential instability in downstream modeling, particularly for algorithms sensitive to feature distribution.

In contrast, the vote_average variable demonstrates a more symmetric and approximately normal distribution, indicating that audience ratings tend to cluster around a central value with less extreme variation. This suggests that while popularity and engagement vary widely, audience satisfaction is

relatively consistent across films. Such behavior reflects the bounded nature of rating systems (e.g., scales from 0 to 10), which naturally limit extreme deviations.

Further analysis of the release year variable reveals a significant temporal bias within the dataset. The distribution is heavily concentrated in the modern era, particularly for films released after the year 2000. This imbalance can be attributed to several factors, including: (1) Increased digital data availability for recent films; (2) Growth of online platforms and user-generated ratings; (3) Higher production and distribution volumes in contemporary cinema. This temporal skew introduces potential bias in analysis and modeling. Models trained on such data may become overfitted to modern trends, reducing their ability to generalize to older films. Additionally, patterns identified in the dataset may reflect recent industry dynamics rather than long-term historical behavior.

The combination of skewed numerical distributions and temporal imbalance highlights the importance of robust preprocessing techniques. Without appropriate handling, these issues may lead to misleading interpretations and reduced model performance. Therefore, subsequent steps such as outlier treatment (Winsorization) and careful feature analysis are essential to ensure a more balanced and representative dataset.

In summary, this analysis confirms that: (1) Popularity and vote-related variables are highly skewed and dominated by extreme values; (2) Rating variables are relatively stable and normally distributed; (3) The dataset exhibits a strong temporal bias toward modern films. These findings form the foundation for subsequent preprocessing and analysis stages, particularly in addressing outliers and ensuring data quality for predictive modeling.

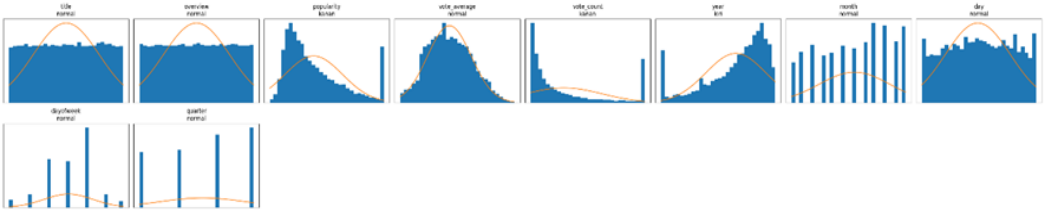

Figure 1. Frequency Distribution of Numerical and Temporal Variables

1. *Outlier Detection and Treatment*

Visual inspection using boxplot analysis confirms the presence of substantial outliers in key numerical variables, particularly popularity and vote_count. A significant number of observations extend beyond the upper whiskers, indicating extreme values that deviate markedly from the majority of the data. This pattern is consistent with the distributional characteristics identified earlier, where a small number of films typically blockbuster productions dominate in terms of audience engagement and visibility.

To complement the visual analysis, a statistical approach using the Interquartile Range (IQR) method was applied to define the acceptable range of values. Based on the calculation: (1) The upper bound for popularity is approximately 9.49; (2) The upper bound for vote_count is approximately 4700. Values exceeding these thresholds are classified as outliers. The large number of such values reinforces the need for careful handling rather than outright removal, as they represent meaningful real-world phenomena rather than noise.

1. *Outlier Treatment Using Winsorization*

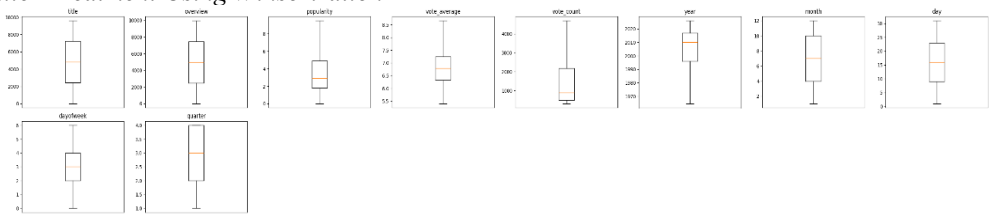

Figure 2. Outlier Identification Through Boxplot Visualization

Instead of applying trimming (which would permanently remove observations), this study adopts a Winsorization (capping) approach. In this method: (1) Values above the upper bound are replaced with the respective threshold values; (2) Lower bound adjustments are applied if necessary, although the primary concern in this dataset lies in upper extremes

This approach is particularly appropriate for the dataset context because: (1) It preserves the presence of influential observations (e.g., highly popular films); (2) It reduces the

disproportionate influence of extreme values on statistical measures; (3) It maintains dataset size and integrity for subsequent analysis.

2. *Post-Treatment Distribution Analysis*

After applying Winsorization, histogram visualizations reveal a more controlled and compact distribution. A noticeable accumulation of values appears at the upper threshold (last bin), reflecting the capping effect. While this introduces a slight artificial clustering, it is an acceptable trade-off for improved statistical stability. More importantly, key statistical indicators show meaningful improvement: (1) The mean shifts closer to the median, indicating reduced skewness; (2) The variance decreases, suggesting a more stable spread of data; (3) The influence of extreme values on overall distribution is significantly minimized.

These changes confirm that the data has become more robust and representative, making it more suitable for downstream analytical tasks, including machine learning modeling.

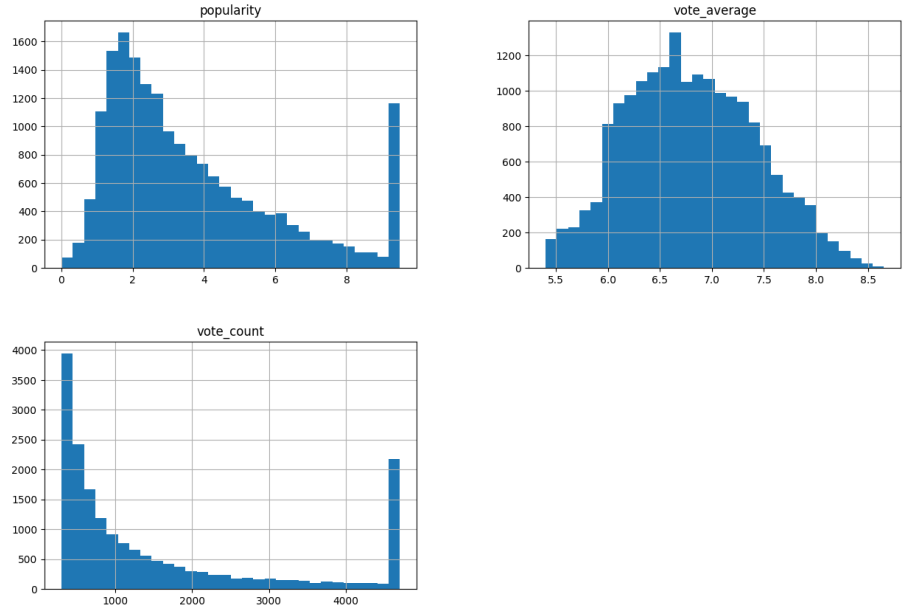

Figure 3. Distribution of Key Variables After Winsorization Application

3. *Discussion and Implications*

The results highlight that outliers in this dataset are not merely anomalies but represent structurally important observations within the film industry ecosystem. Therefore, completely removing them could lead to loss of critical information, especially regarding high-performing films. By applying Winsorization, this study achieves a balance between: (1) Data integrity (preserving observations); (3) Statistical robustness (reducing extreme influence).

However, it is important to note that Winsorization may introduce distributional distortion at boundary values, particularly when a large proportion of data is capped. Future work could explore alternative robust techniques such as: (1) Log transformation; (2) Robust scaling; (3) Quantile transformation. In summary, the outlier handling strategy successfully improves the dataset by: (1) Reducing extreme skewness; (2) Stabilizing statistical properties; (3) Preserving meaningful high-value observations.

This preprocessing step plays a critical role in ensuring that subsequent analyses and predictive models are both reliable and generalizable.

B. Correlation and Multicollinearity Analysis

The relationships among variables were evaluated using the Pearson correlation matrix to quantify the strength and direction of linear associations between features. This analysis provides critical insights into how variables interact and helps identify potential redundancy that may affect downstream modeling.

1. Correlation Analysis Results

Based on Figure 4, a moderately strong positive correlation ($r \approx 0.62$) is observed between popularity and vote_count. This finding aligns with intuitive expectations, as films that receive higher visibility and audience attention tend to accumulate more votes. From a data perspective,

this relationship reflects user engagement dynamics, where popularity acts as a driver for audience participation. However, the correlation is not perfect, indicating that popularity and vote count capture related but distinct dimensions: (1) Popularity may be influenced by marketing, trends, or platform exposure; (2) Vote count reflects actual user interaction and participation.

This distinction is important because it suggests that both variables still contribute unique information and should not be immediately removed without further evaluation. In contrast, the vote_average variable shows very weak correlation with both popularity and vote_count. This implies that audience ratings are relatively independent of how widely a movie is viewed or discussed. In other words, a film can be highly popular but not necessarily highly rated, and vice versa. This finding highlights the difference between quantity (engagement) and quality (perceived value) in movie evaluation.

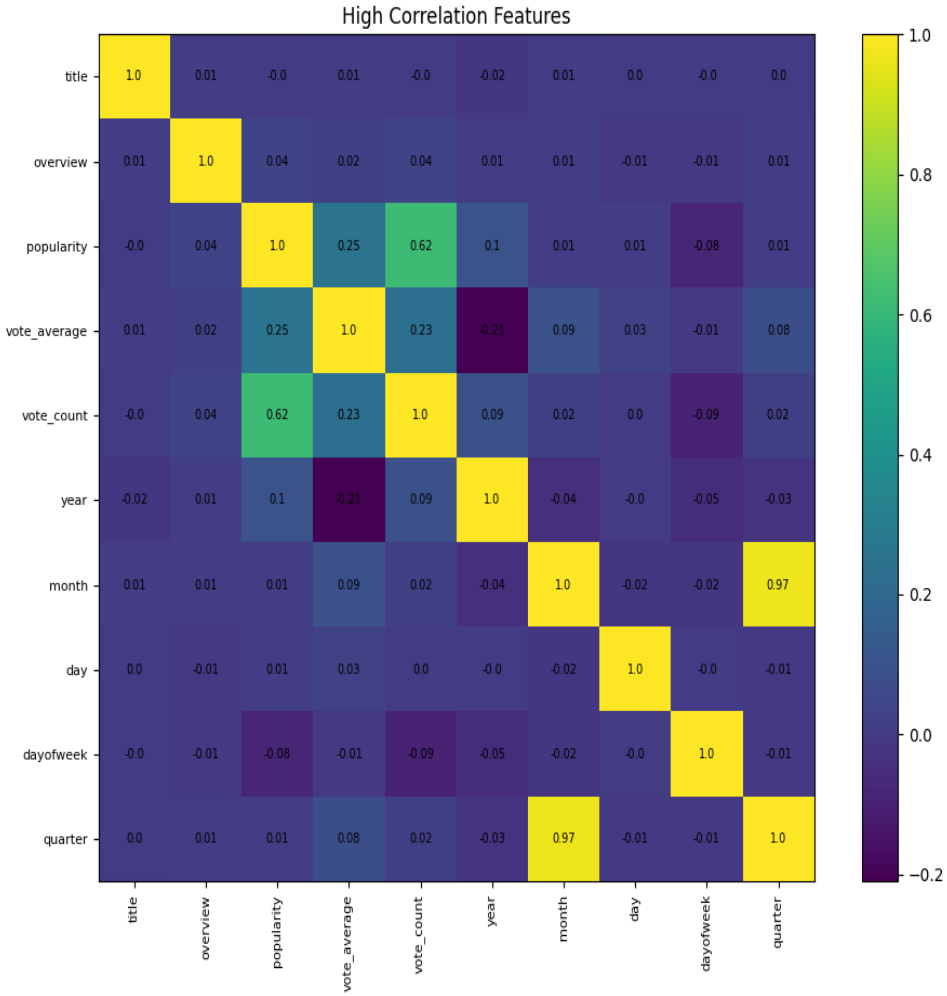

Figure 4. Correlation Matrix of Variables

2. Multicollinearity Detection

A key finding in this analysis is the presence of near-perfect correlation ($r \approx 0.97$) between the month and quarter variables. This indicates a case of severe multicollinearity, where one variable can almost entirely explain the other. This issue arises because: (1) Quarter is a direct transformation of month; (2) Both variables encode the same temporal information at different granularities. Such redundancy can have negative implications, particularly in predictive modeling: (1) It can inflate variance in model coefficients; (2) It reduces model interpretability; (3) It may lead to instability in regression-based algorithms.

3. Discussion and Implications

The correlation analysis reveals three important insights:

a. Meaningful Relationships Exist

The moderate correlation between popularity and vote count confirms expected behavioral patterns in audience engagement.

b. Feature Independence is Preserved in Some Variables

The weak correlation of vote_average suggests that it provides independent information, making it a valuable feature for modeling.

c. Redundant Features Must Be Addressed

The strong multicollinearity between month and quarter highlights the need for feature selection.

To address multicollinearity, several strategies can be considered: (1) Removing one of the correlated variables (e.g., retaining month and dropping quarter); (2) Applying dimensionality reduction techniques; (3) Using models that are less sensitive to multicollinearity (e.g., tree-based models).

Overall, the correlation and multicollinearity analysis ensures that the dataset is statistically sound and free from redundant information. By identifying both meaningful relationships and problematic dependencies, this step plays a crucial role in improving: (1) Model stability; (2) Interpretability; (3) Predictive performance. These findings serve as a foundation for feature selection and further modeling, ensuring that only relevant and non-redundant variables are utilized in subsequent stages.

IV. CONCLUSION

This study conducted a comprehensive Exploratory Data Analysis (EDA) on the Top Rated Movies from TMDb dataset using a robust preprocessing approach based on Winsorization. The analysis revealed several important findings related to data distribution, outlier characteristics, and relationships among variables. First, the distribution analysis showed that key variables such as popularity and vote_count exhibit highly right-skewed distributions, indicating the presence of extreme values dominated by a small number of blockbuster films. In contrast, vote_average demonstrated a more stable and near-normal distribution. Additionally, a significant temporal bias was identified, with the dataset heavily concentrated on modern films released after the year 2000. Second, the outlier detection process using the IQR method confirmed the existence of substantial extreme values. To address this issue without losing critical information, the study applied the Winsorization technique. The results showed that this approach successfully reduced skewness, stabilized variance, and improved the overall representativeness of the dataset, while preserving important observations related to high-performing films. Third, correlation analysis revealed a moderately strong relationship between popularity and vote_count, indicating consistent audience engagement patterns. Meanwhile, vote_average showed weak correlation with other variables, suggesting that rating quality is relatively independent of popularity. The study also identified severe multicollinearity between month and quarter, highlighting the need for feature selection to avoid redundancy.

Overall, this research demonstrates that a structured and robust EDA process—combining statistical techniques such as IQR, Winsorization, and correlation analysis—plays a critical role in improving data quality. The resulting dataset becomes more balanced, stable, and suitable for further analytical modeling. In conclusion, the application of Winsorization proves to be an effective strategy for handling outliers in real-world datasets characterized by extreme distributions. The findings of this study provide a strong foundation for future work in predictive modeling, recommendation systems, and data-driven decision-making within the film industry.

V. REFERENCES

[1] Jiawei Han, M. Kamber, & J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2012.

[2] Peter J. Rousseeuw & A. M. Leroy, Robust Regression and Outlier Detection. Wiley, 1987.

[3] John W. Tukey, Exploratory Data Analysis. Addison-Wesley, 1977.

[4] Rand R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, 4th ed. Academic Press, 2017.

[5] Trevor Hastie, Robert Tibshirani, & Jerome Friedman, The Elements of Statistical Learning. Springer, 2009.

[6] John W. Tukey, Exploratory Data Analysis. Addison-Wesley, 1977.

[7] Gareth James et al., An Introduction to Statistical Learning. Springer, 2013.

[8] Douglas C. Montgomery, E. A. Peck, & G. G. Vining, Introduction to Linear Regression

Analysis. Wiley, 2012.
[9] Andrew Ng, "Machine Learning Yearning," 2018.
[10] TMDb, TMDb Dataset Documentation, 2026.
[11] I. Goodfellow, Y. Bengio, & A. Courville, *Deep Learning*. MIT Press, 2016.
[12] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. O'Reilly, 2022.
[13] Hadley Wickham & Garrett Grolemund, *R for Data Science*. O'Reilly, 2017.
[14] Max Kuhn & Kjell Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press, 2019.
[15] J. Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. Machine Learning Mastery, 2020.
[16] Hidayat, I., Tolago, A. I., Dako, R. D. R., & Ilham, J. (2023). Analisis Data Eksploratif Capaian Indikator Kinerja Utama 3 Fakultas Teknik. Jambura Journal of Electrical and Electronics Engineering, 5(2), 185-191.