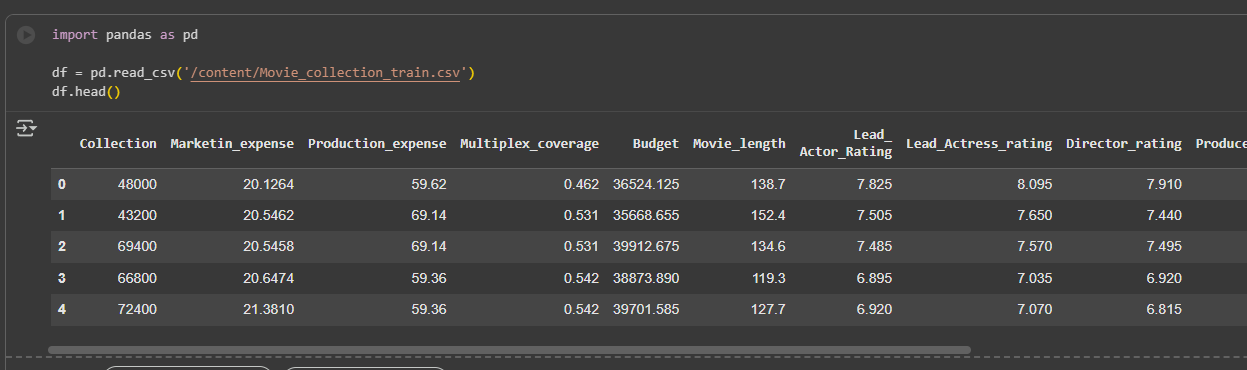
Movie Collection Success Analysis  
  
**Introduction**

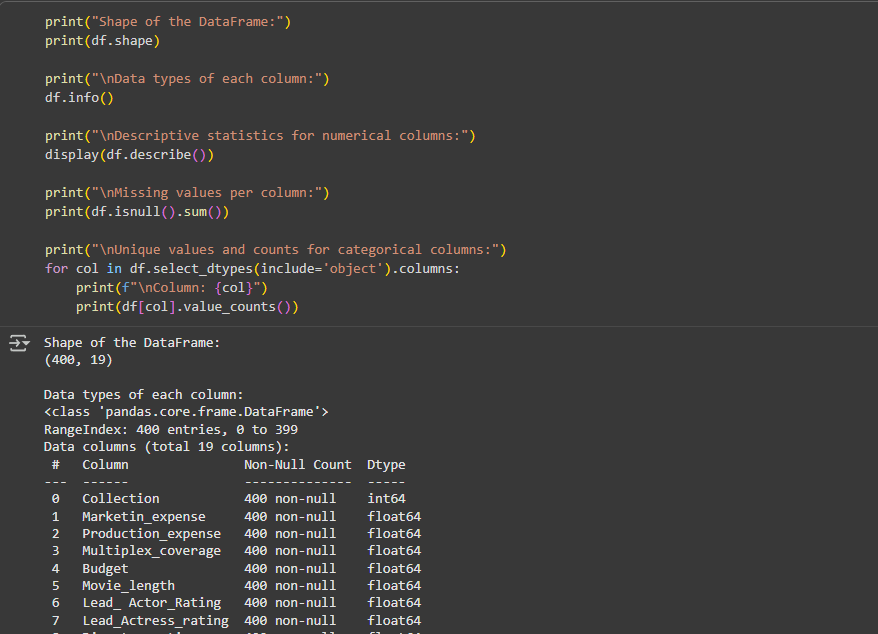
The movie industry is highly competitive, and predicting the commercial success of a film is critical for production houses and investors. This project analyzes the dataset *Movie\_collection\_train.csv* to identify factors influencing success and build a predictive model. The focus is on applying Logistic Regression in combination with Linear Discriminant Analysis (LDA) for dimensionality reduction and classification.

**1. Data Loading and Exploration**

The dataset consisted of **400 rows and 19 columns**, with both numerical and categorical variables. Key findings from the initial exploration:

* **Missing Values**: 8 missing values were found in the Time\_taken column.
* **Redundant Feature**: The MPAA\_film\_rating column contained only a single unique value and was dropped.
* **Categorical Features**: Genre and 3D\_available were identified for encoding.
* **Target Creation**: A binary target variable, Collection\_Success, was derived. Movies above the median Collection value were labeled as successful (1), while others were labeled unsuccessful (0).



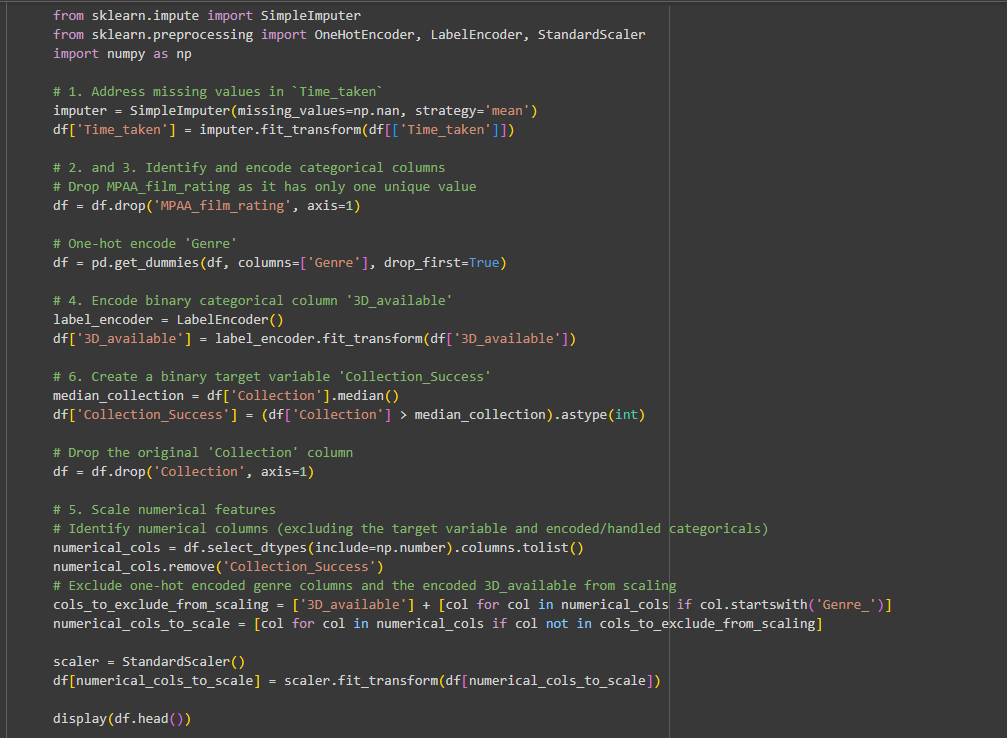


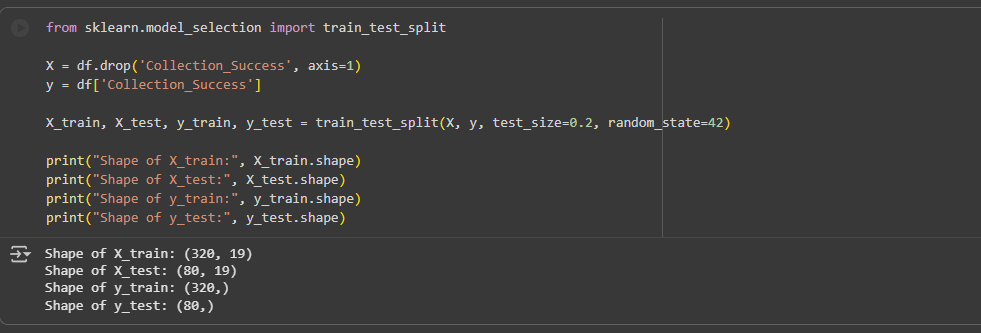
**2. Data Preprocessing**

Steps undertaken:

* Missing values in Time\_taken imputed with the **mean**.
* Genre was **one-hot encoded**, while 3D\_available was **label encoded**.
* Original Collection column was dropped after creating the target variable.
* Numerical features were scaled using **StandardScaler**.

This ensured consistency and comparability across features.

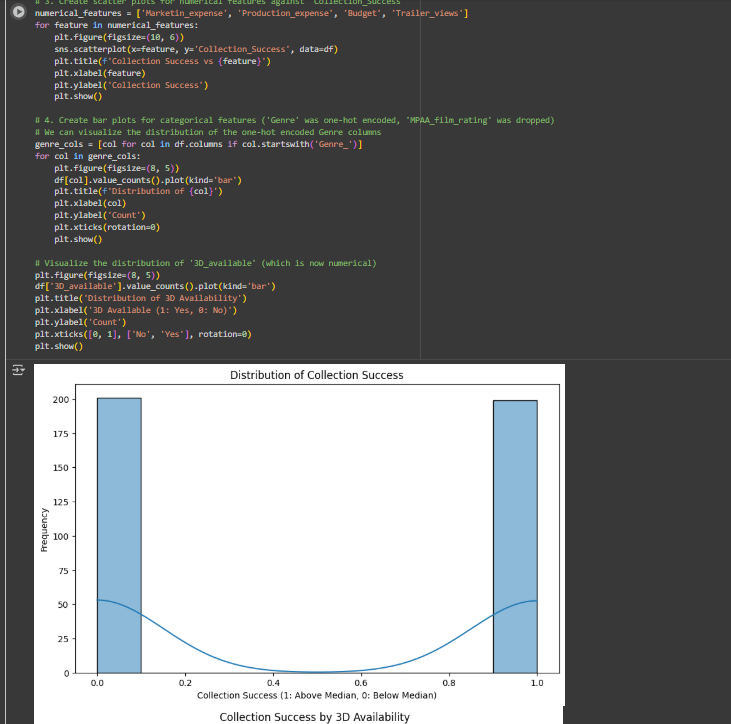




**3. Data Visualization**

Exploratory data analysis highlighted several insights:

* **Target Distribution**: The histogram of Collection\_Success showed a relatively balanced split.
* **Box and Scatter Plots**: Features like Marketing\_expense, Production\_expense, Budget, and Trailer\_views displayed clear differences between successful and unsuccessful movies, though some overlap remained.
* **Categorical Features**: Bar plots revealed distinct distributions for Genre and 3D\_available, with 3D releases showing a higher association with success.



**4. Linear Discriminant Analysis (LDA)**

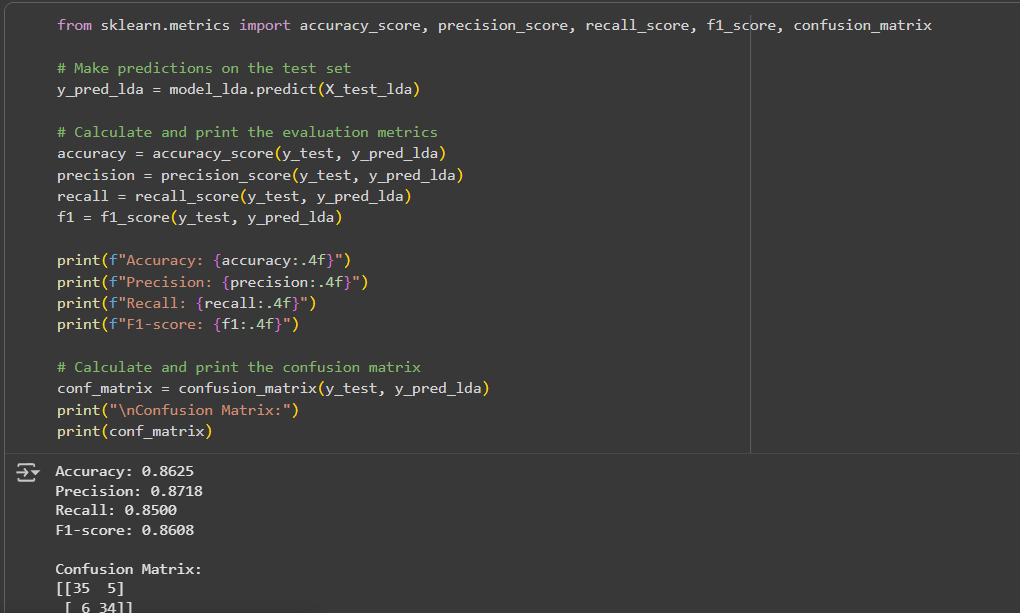
LDA was applied to reduce the feature space. Since this was a binary classification task, features were projected onto a **single discriminant axis**. This transformation maximized class separability by capturing the most informative linear combinations of features.

**5. Model Building and Evaluation**

A **Logistic Regression model** was trained on the LDA-transformed training data. Performance was assessed on the test set:

* **Accuracy**: 0.8625
* **Precision**: 0.8718
* **Recall**: 0.8500
* **F1-Score**: 0.8608

The **confusion matrix** confirmed a strong balance between true positives and true negatives, with minimal misclassifications.



**6. Interpretation of Results**

The logistic regression coefficient for the LDA component was **+2.04**, indicating a strong positive link between the discriminant axis and success probability. This suggests that LDA effectively condensed the predictive signals from multiple features.

From earlier modeling stages, **Trailer\_views, Budget, and 3D\_available** emerged as significant contributors. Their combined representation in the LDA component reinforces their predictive importance.

**Conclusion**

This project demonstrates that **Logistic Regression, enhanced with LDA**, provides an effective approach for predicting movie collection success. With an F1-score of 0.8608, the model balances precision and recall, making it practical for real-world applications such as forecasting film revenues.

**Key Insights:**

* Trailer\_views, Budget, and 3D\_available are influential features.
* LDA helps condense feature space into a highly discriminative axis.
* Logistic Regression offers interpretability and competitive performance.