**YouTube Earnings Estimation: Data Processing and Modeling**

**Introduction**

The goal of this analysis was to estimate the potential earnings of YouTube channels based on available metrics such as views, likes, comments, and video duration. The data was collected from a set of YouTube channels, and a machine learning model was applied to predict earnings. This report details the process of data collection, preprocessing, model selection, and the results of the analysis.

**Data Collection**

The data used in this analysis was sourced from public YouTube statistics. It includes the following metrics for each video:

* **subscriberCount**: The number of subscribers the channel has.
* **viewCount**: The total number of views the channel has received.
* **videoCount**: The total number of videos uploaded by the channel.
* **Views**: The number of views for the specific video.
* **Likes**: The number of likes the video received.
* **Comments**: The number of comments on the video.
* **Duration (seconds)**: The duration of the video in seconds.
* **Is Live**: A boolean indicating whether the video is a live stream.
* **Published At**: The date and time the video was published.

This data was exported from YouTube's API and saved in CSV format for analysis.

**Data Preprocessing**

Before modeling, the data underwent several preprocessing steps to ensure its quality and suitability for machine learning:

1. **Date Parsing**: The Published At column was in ISO 8601 format (e.g., 2024-08-17T16:20:07Z). This format includes a 'T' between the date and time and a 'Z' indicating UTC time. The dates were parsed and converted into a datetime format that could be used for analysis.
2. **Handling Missing Values**: Missing values were imputed using the mean of the respective columns to avoid any disruption in model training. This was applied only to numeric columns like Views, Likes, and Comments.
3. **Feature Conversion**: The Is Live column, which was initially a boolean, was converted into an integer (1 for True, 0 for False) for compatibility with the machine learning model.

**Model Selection**

A **Linear Regression** model was selected for this analysis due to its simplicity and effectiveness in predicting continuous outcomes like earnings. Linear Regression is a widely used algorithm in machine learning for tasks involving the prediction of a dependent variable (in this case, estimated earnings) based on one or more independent variables (such as views, likes, comments, etc.).

**Model Training and Evaluation**

The dataset was divided into training and testing sets, with 80% of the data used for training the model and 20% reserved for testing its accuracy. The model was trained to predict the Estimated Earnings based on the provided features.

* **Mean Absolute Error (MAE)**: This metric was used to evaluate the model's performance. It represents the average absolute difference between the predicted and actual values. A lower MAE indicates better predictive accuracy.
* **R-squared (R²)**: This statistical measure was also used to evaluate the model, representing the proportion of variance in the dependent variable that is predictable from the independent variables. An R² value closer to 1 indicates a better fit of the model.

**Earnings Estimation**

To estimate the earnings for each video, a base **CPM (Cost Per Mille)** of $2 was assumed. CPM represents the revenue earned per 1,000 ad impressions. The formula used was:

Estimated Earnings=Views×CPM1000\text{Estimated Earnings} = \frac{\text{Views} \times \text{CPM}}{1000}Estimated Earnings=1000Views×CPM​

Given the conservative CPM value, this estimation provides a basic idea of potential earnings based on the video's view count.

**Monthly Earnings Calculation**

To provide a broader view of potential earnings over time, monthly earnings were also calculated. The videos were grouped by their publication month, and the views were summed to estimate total monthly earnings using the same CPM assumption.

**Results**

The results of the model included:

* **Predicted Earnings for Each Video**: These were compared to the actual earnings to assess model accuracy.
* **Feature Importance**: This analysis revealed which features (e.g., views, likes, comments) were most influential in predicting earnings.
* **Estimated Monthly Earnings**: This provides an overview of potential revenue generated on a monthly basis, helping to understand the channel's financial performance over time.

**Conclusion**

This analysis provides a structured approach to estimating YouTube earnings based on publicly available metrics. By applying a Linear Regression model and carefully preprocessing the data, we achieved a reasonable estimation of potential revenue. While the model uses a conservative CPM value, it can be adjusted based on more specific knowledge of the channel's audience and advertiser interest.

This approach can be extended to other YouTube channels and can be refined further by incorporating more detailed data, such as audience demographics and engagement metrics, which could improve the accuracy of the earnings predictions.