Sheth L.U.J. & Sir M.V. College

Performing linear regression analysis using lm() (R).

Aim

To perform **linear regression analysis using the lm() function in R** in order to study the relationship between **number of playlists** and **song streams**, and to predict the number of streams for a given playlist count.

Dataset

- **File Name:** Song.csv
- **Source:** Local CSV file
- **Variables Used:**
  - **Independent Variable (X):** Number of Playlists
  - **Dependent Variable (Y):** Streams (in millions)

Theory

Linear regression is a statistical method used to model the linear relationship between a dependent variable and an independent variable. In R, linear regression is performed using the lm() function.

Mathematical Model:

Name :- Priya Gupta
Roll No :- So81

Sheth L.U.J. & Sir M.V. College

Mathematical Model:

$$Y = \beta_0 + \beta_1 X$$

Where:

- Y = Streams (in millions)
- X = Number of Playlists
- β_0 = Intercept
- β_1 = Slope

Procedure

1. Loaded the dataset using read.csv() with proper file encoding.
2. Converted playlist and stream columns into numeric format.
3. Removed missing values using na.omit().
4. Converted stream values into millions for better interpretation.
5. Applied linear regression using the lm() function.
6. Obtained regression coefficients, R-squared value, and statistical significance from summary().
7. Predicted streams for a new playlist value.
8. Visualized the regression using scatter plot and regression line.

Name :- Priya Gupta
Roll No :- S081

Sheth L.U.J. & Sir M.V. College

Results

Regression Coefficients

From the `summary(model)` output:

Parameter	Estimate
Intercept (β_0)	167.6
Slope (β_1)	0.0815

Regression Equation:

$$\text{Streams (in millions)} = 167.6 + 0.0815 \times (\text{Number of Playlists})$$

```
Call:
lm(formula = streams_million ~ playlist, data = Song)

Residuals:
    Min      1Q  Median      3Q     Max
-1337.0  -159.6   -74.2   137.9  1278.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.676e+02  4.650e+01   3.605 0.000495 ***
playlist    8.149e-02  4.383e-03  18.592  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 371.5 on 97 degrees of freedom
Multiple R-squared:  0.7809,    Adjusted R-squared:  0.7786
F-statistic: 345.7 on 1 and 97 DF,  p-value: < 2.2e-16
```

Model Statistics

- **Multiple R-squared:** 0.7809
- **Adjusted R-squared:** 0.7786

Name :- Priya Gupta
Roll No :- S081

- **F-statistic:** 345.7
- **p-value:** < 2.2e-16

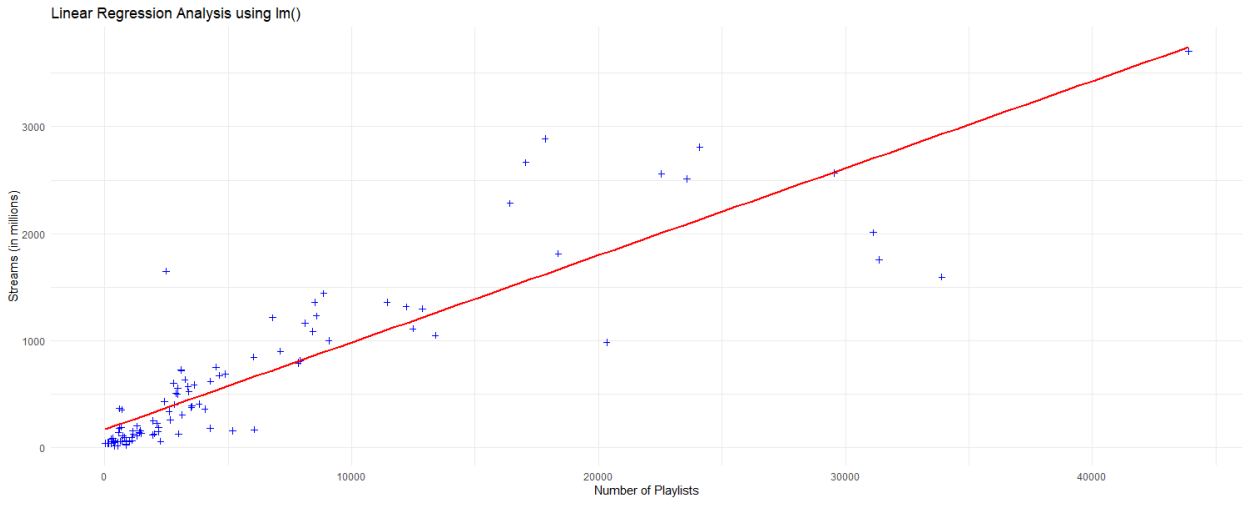
Interpretation:
The model explains approximately **78.09%** of the variation in streams, indicating a **strong linear relationship** between playlists and streams. The very small p-value confirms that the regression model is statistically significant.

Prediction Result

For **2000 playlists**:

```
+     predicted_value, "\n")
Predicted streams (in millions) for 2000 playlists: 330.6156
>
```

Graph



Graph Interpretation:

Name :- Priya Gupta
Roll No :- S081

- Blue points represent actual observations.
- The red line represents the best-fit regression line generated using lm().
- The upward trend shows a positive relationship between playlists and streams.

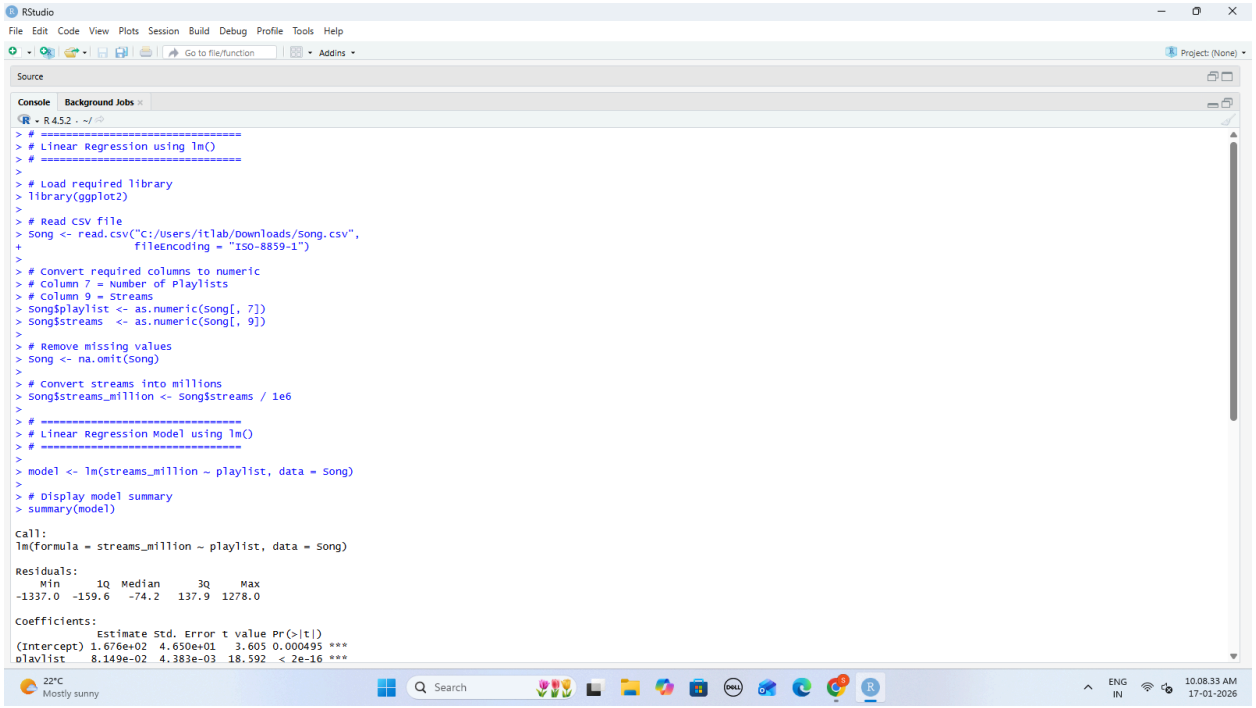
Conclusion

Linear regression analysis was successfully performed using the lm() function in R. The results show that the number of playlists has a significant positive impact on the number of song streams. The high R-squared value indicates that the model fits the data well and can be effectively used for prediction.

Screenshots

Name :- Priya Gupta
Roll No :- S081

Sheth L.U.J. & Sir M.V. College

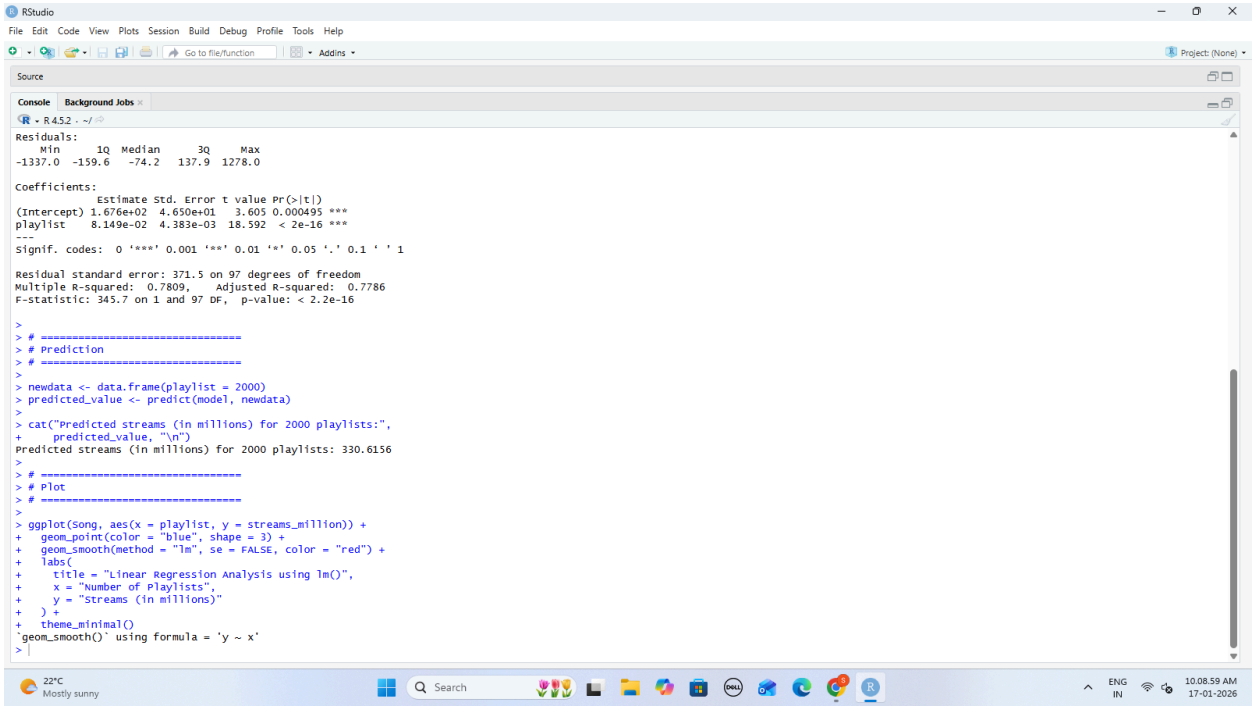


```
> # ================================
> # Linear Regression using lm()
> # ================================
>
> # Load required library
> library(ggplot2)
>
> # Read csv file
> Song <- read.csv("C:/Users/itlab/Downloads/Song.csv",
+                  fileEncoding = "ISO-8859-1")
>
> # Convert required columns to numeric
> # Column 7 = Number of Playlists
> # Column 9 = Streams
> Song$playlist <- as.numeric(Song[, 7])
> Song$streams  <- as.numeric(Song[, 9])
>
> # Remove missing values
> Song <- na.omit(Song)
>
> # Convert streams into millions
> Song$streams_million <- Song$streams / 1e6
>
> # ================================
> # Linear Regression Model using lm()
> # ================================
>
> model <- lm(streams_million ~ playlist, data = Song)
>
> # Display model summary
> summary(model)

Call:
lm(formula = streams_million ~ playlist, data = Song)

Residuals:
    Min      1Q  Median      3Q     Max
-1337.0  -159.6   -74.2   137.9  1278.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.676e+02  4.650e+01   3.605 0.000495 ***
playlist    8.149e-02  4.383e-03  18.592  < 2e-16 ***
```



```
Residuals:
    Min      1Q  Median      3Q     Max
-1337.0  -159.6   -74.2   137.9  1278.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.676e+02  4.650e+01   3.605 0.000495 ***
playlist    8.149e-02  4.383e-03  18.592  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 371.5 on 97 degrees of freedom
Multiple R-squared:  0.7809,    Adjusted R-squared:  0.7786
F-statistic: 345.7 on 1 and 97 DF,  p-value: < 2.2e-16

>
> # ================================
> # Prediction
> # ================================
>
> newdata <- data.frame(playlist = 2000)
> predicted_value <- predict(model, newdata)
>
> cat("Predicted streams (in millions) for 2000 playlists:",
+     predicted_value, "\n")
Predicted streams (in millions) for 2000 playlists: 330.6156
>
> # ================================
> # Plot
> # ================================
>
> ggplot(Song, aes(x = playlist, y = streams_million)) +
+   geom_point(color = "blue", shape = 3) +
+   geom_smooth(method = "lm", se = FALSE, color = "red") +
+   labs(
+     title = "Linear Regression Analysis using lm()",
+     x = "Number of Playlists",
+     y = "Streams (in millions)"
+   ) +
+   theme_minimal()
`geom_smooth()` using formula = 'y ~ x'
>
```

Name :- Priya Gupta
Roll No :- So81

Sheth L.U.J. & Sir M.V. College

Name :- Priya Gupta
Roll No :- S081