VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A2b: IPL Performance vs Salary

GAURI VINOD NAIR

V01110160

Date of Submission: 23-06-2024

TABLE OF CONTENTS

INTRODUCTION

In this study, we delve into the dynamic relationship between player performance and compensation in the context of the Indian Premier League (IPL). Utilizing a comprehensive dataset, "IPL_ball_by_ball_updated till 2024.csv," we aim to rigorously analyze how player salaries correspond to their on-field performance over the last three years. This dataset encompasses a wealth of attributes including match details, batting and bowling statistics, player contributions, and outcomes.

The IPL serves as an ideal platform to explore these relationships due to its high-profile nature and the availability of detailed statistical records. By employing regression analysis techniques, we seek to establish quantifiable links between a player's performance metrics—such as runs scored, wickets taken, and fielding contributions—and the financial compensation they receive.

Our objectives are twofold: first, to uncover statistically significant correlations between player performance metrics and salary levels; and second, to interpret these findings in a manner that provides actionable insights for team management, player agents, and league stakeholders. This analysis not only sheds light on the economic dynamics within the IPL but also contributes to a deeper understanding of how performance metrics influence financial incentives and player valuation in professional cricket.

The insights gained from this study are expected to inform strategic decisions related to player recruitment, contract negotiations, and team composition, thereby optimizing team performance and enhancing the overall competitiveness of the IPL. By leveraging empirical data and advanced analytical techniques, we aim to contribute valuable insights that resonate across the realms of sports management, economics, and performance analytics.

OBJECTIVE

1. Establish Relationship Through Regression Analysis
   - Utilize regression analysis to quantify and understand how various performance metrics (such as runs scored, wickets taken, fielding contributions) relate to player salaries in the IPL.
   - Identify which performance metrics have the strongest association with player compensation.
2. Explore Trends Over Three Years
   - Analyze trends in the relationship between player salaries and performance metrics over the last three IPL seasons.
   - Determine if there are consistent patterns or if relationships have evolved over time.
3. Identify Significant Factors
   - Identify other factors beyond basic performance metrics (e.g., match conditions, player experience, team success) that may influence player salaries.
   - Assess their impact on salary determination alongside direct performance metrics.

BUSINESS SIGNIFICANCE

Analyzing the relationship between player performance and payment in the Indian Premier League (IPL) through regression analysis holds significant business implications across several key areas. Firstly, understanding how performance metrics such as runs scored, wickets taken, and fielding contributions correlate with player salaries enables IPL team management to optimize their resource allocation. By aligning player salaries with their on-field contributions, teams can make more informed decisions regarding player recruitment, retention, and contract negotiations. This approach not only enhances team performance but also maximizes return on investment (ROI), ensuring that resources are allocated efficiently to achieve competitive success.

Moreover, this analysis facilitates strategic team composition by identifying players whose performance metrics justify their compensation. Teams can assemble squads that balance performance excellence with financial prudence, thereby enhancing their chances of success in the highly competitive IPL environment. This strategic alignment of player salaries with performance metrics also supports better negotiation strategies for player agents and representatives. Armed with empirical evidence of their clients' contributions on the field, agents can advocate more effectively for fair compensation during contract negotiations, promoting a transparent and merit-based approach to player valuation.

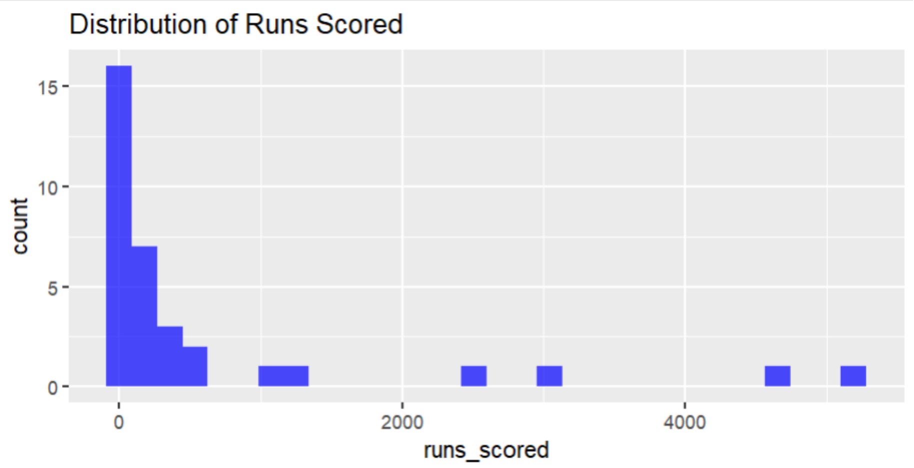
Beyond team management and player negotiations, understanding the relationship between performance and payment enhances fan engagement and sponsorship value. High-performing players often attract greater fan interest and support, which translates into increased attendance, merchandise sales, and sponsorship opportunities for IPL franchises. By investing in players based on objective performance metrics, teams can foster stronger fan loyalty and support, thereby enhancing the overall commercial viability and marketability of the league.

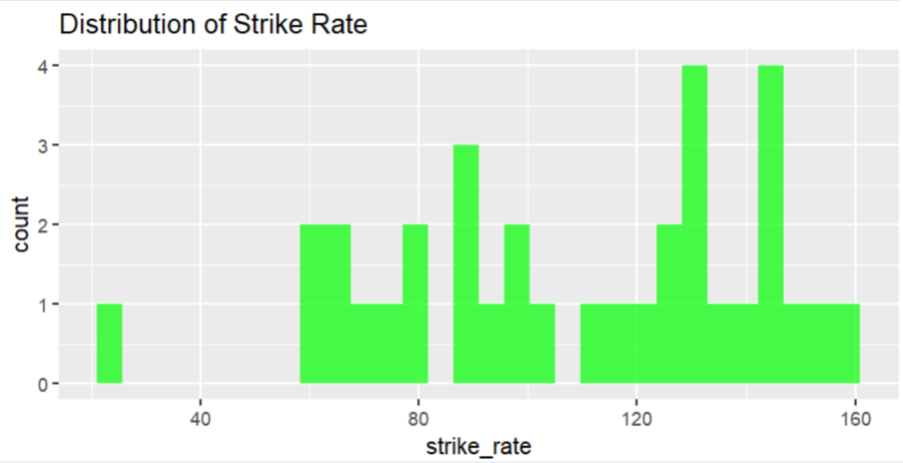
Furthermore, emphasizing data-driven decision-making in player valuation promotes fairness and transparency within the IPL ecosystem. It reduces the influence of subjective evaluations and biases, ensuring that players are compensated based on their actual contributions on the field rather than perceived value alone. This approach not only enhances the credibility of player contracts but also contributes to a more competitive and equitable environment within the league.

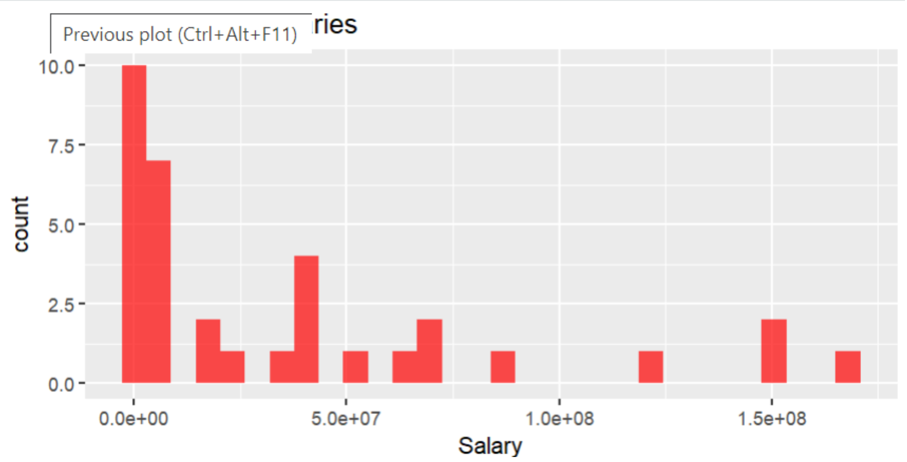
REAULTS AND INTERPRETAIONS

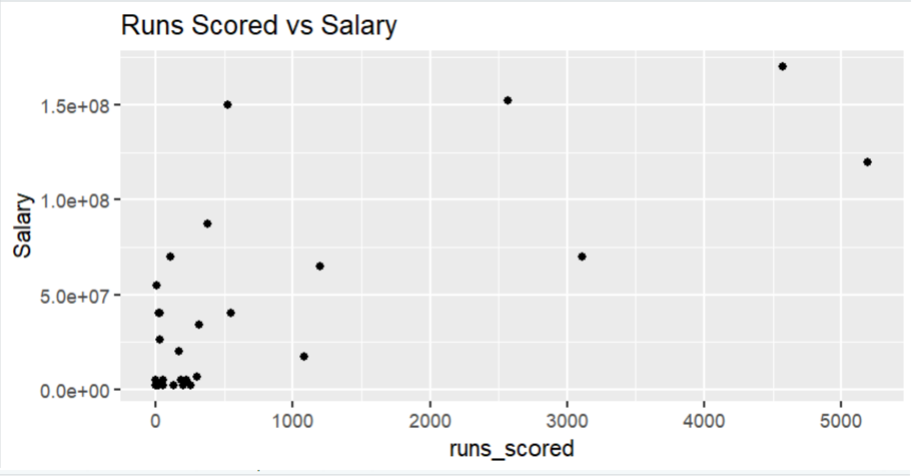
1. Exploratory Data Analysis

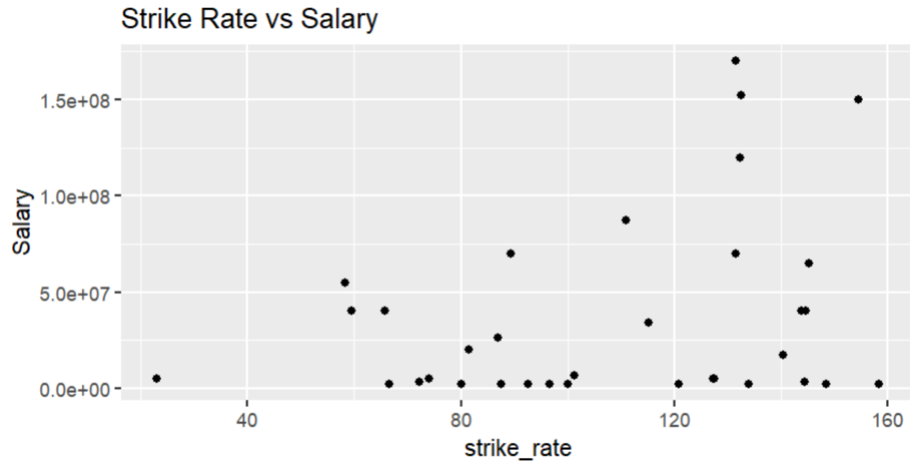
R



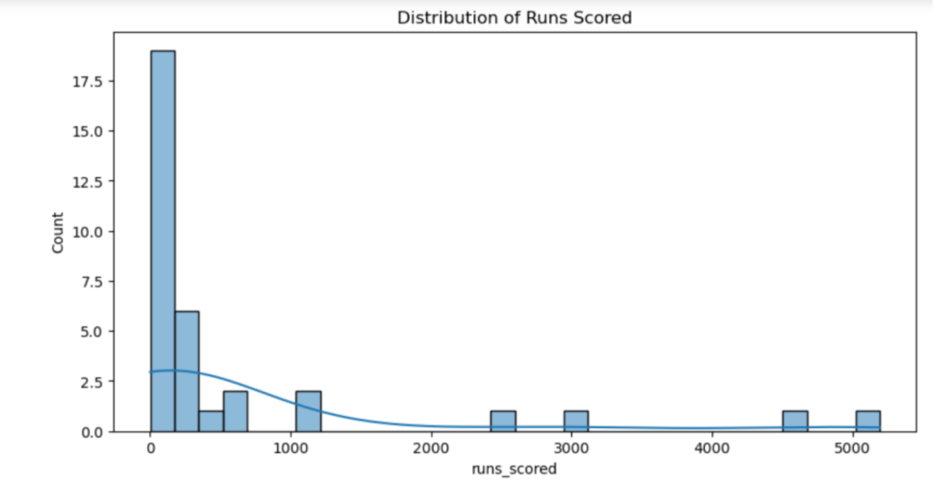


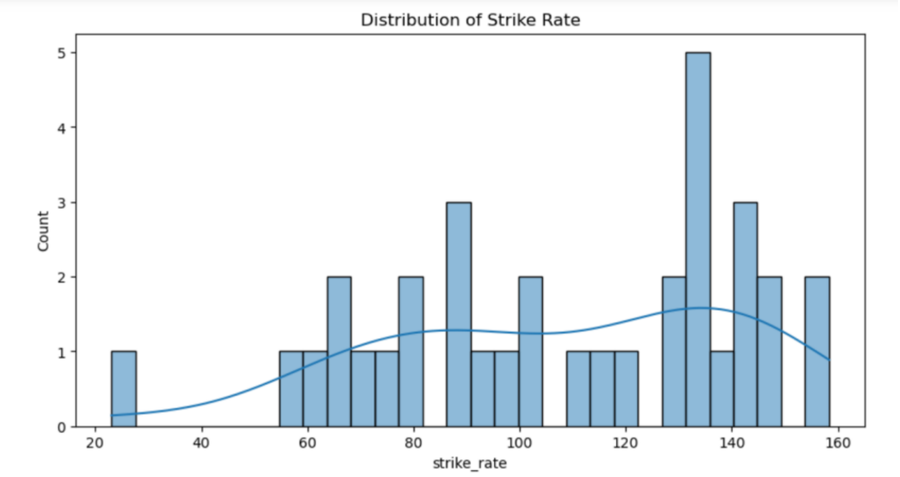


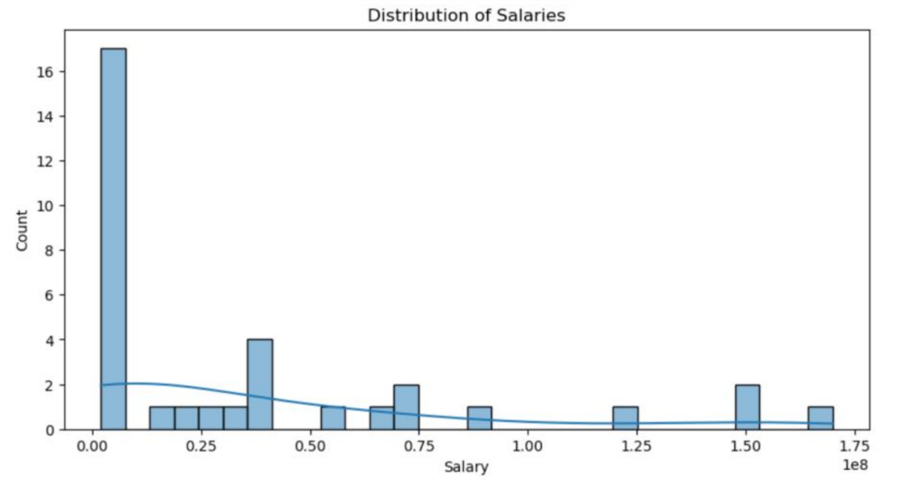

Runs Scored vs Salary
Salary
1.5e+08
1.0e+08
5.0e+07
0.0e+00
0
1000
2000
3000
4000
5000
runs_scored

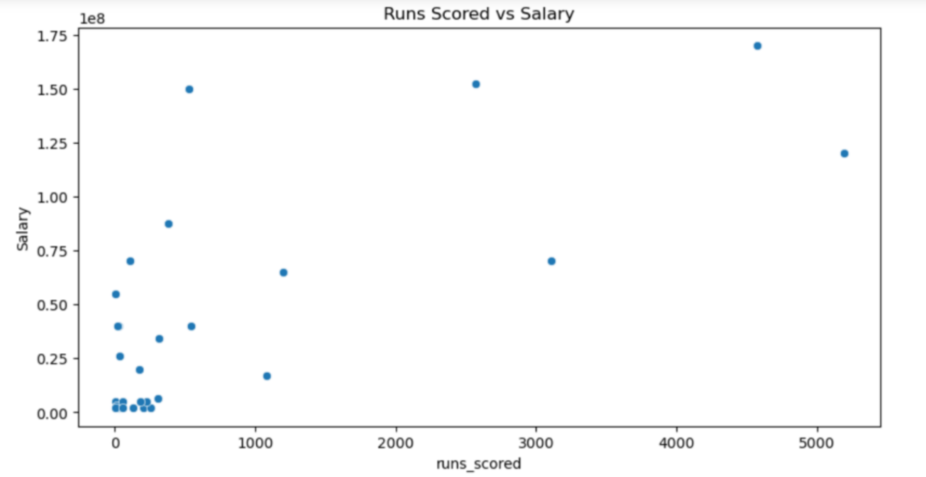

Strike Rate vs Salary
Salary
1.5e+08
1.0e+08
5.0e+07
0.0e+00
40
80
120
160
strike_rate

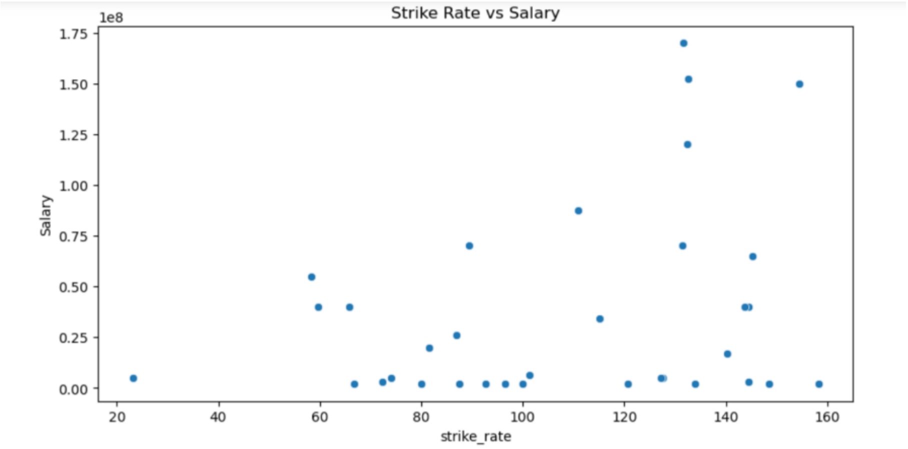
# Python


Distribution of Runs Scored
Count
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0
1000
2000
3000
4000
5000
runs_scored


Distribution of Strike Rate
Count
0
1
2
3
4
5
20
40
60
80
100
120
140
160
strike_rate

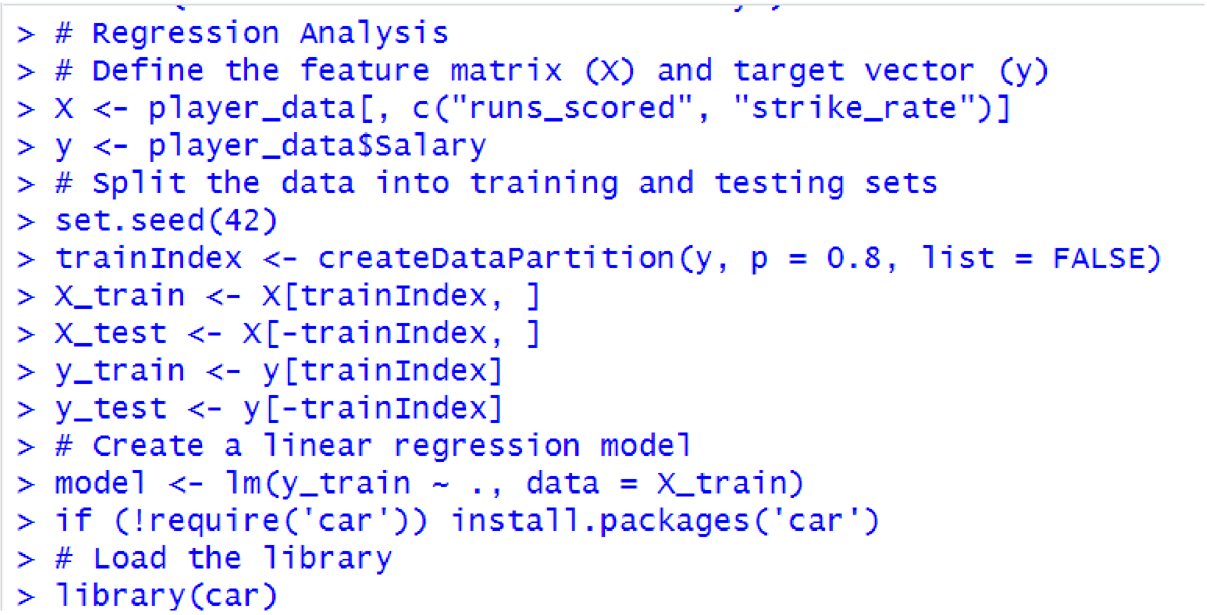

Distribution of Salaries
Count
0
2
4
6
8
10
12
14
16
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e8
Salary


Runs Scored vs Salary
1e8
Salary
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
1000
2000
3000
4000
5000
runs_scored

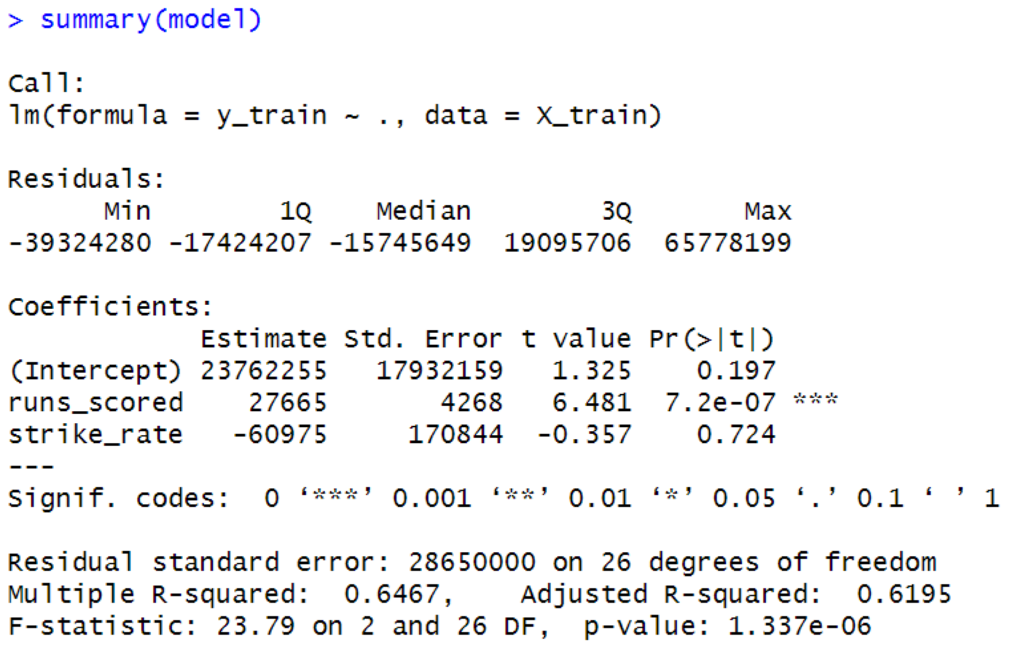


2. Regression Analysis

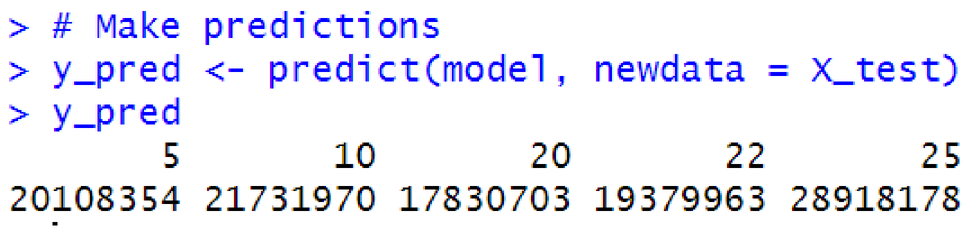
R



```
> # Regression Analysis
> # Define the feature matrix (X) and target vector (y)
> X <- player_data[, c("runs_scored", "strike_rate")]
> y <- player_data$Salary
> # Split the data into training and testing sets
> set.seed(42)
> trainIndex <- createDataPartition(y, p = 0.8, list = FALSE)
> X_train <- X[trainIndex, ]
> X_test <- X[-trainIndex, ]
> y_train <- y[trainIndex]
> y_test <- y[-trainIndex]
> # Create a linear regression model
> model <- lm(y_train ~ ., data = X_train)
> if (!require('car')) install.packages('car')
> # Load the library
> library(car)
```

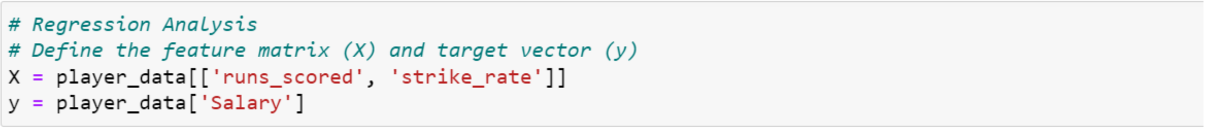


```
> summary(model)

Call:
lm(formula = y_train ~ ., data = X_train)

Residuals:
      Min        1Q    Median        3Q       Max 
-39324280 -17424207 -15745649  19095706  65778199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23762255   17932159   1.325    0.197    
runs_scored    27665       4268   6.481  7.2e-07 ***
strike_rate   -60975     170844  -0.357    0.724    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28650000 on 26 degrees of freedom
Multiple R-squared:  0.6467,	Adjusted R-squared:  0.6195 
F-statistic: 23.79 on 2 and 26 DF,  p-value: 1.337e-06
```

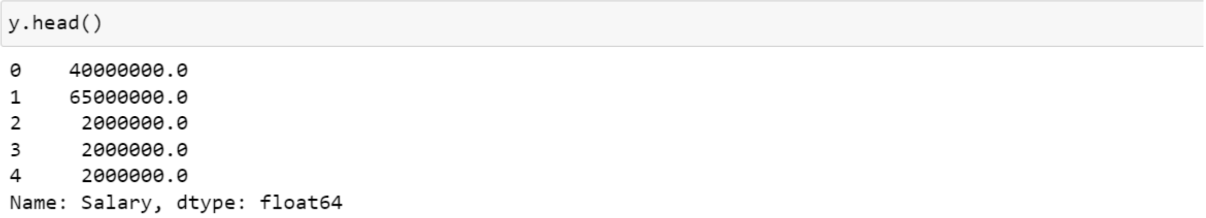


```
> # Make predictions
> y_pred <- predict(model, newdata = X_test)
> y_pred
       5       10       20       22       25 
20108354 21731970 17830703 19379963 28918178 
```

# Python

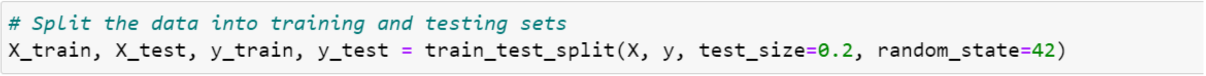


```
# Regression Analysis
# Define the feature matrix (X) and target vector (y)
X = player_data[['runs_scored', 'strike_rate']]
y = player_data['Salary']
```

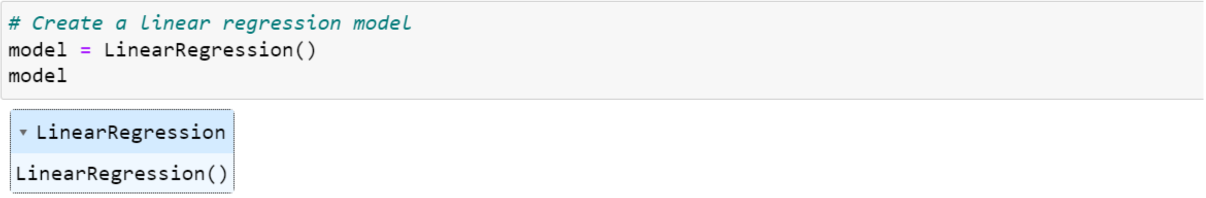


```
y.head()
```

```
0    40000000.0
1    65000000.0
2     2000000.0
3     2000000.0
4     2000000.0
Name: Salary, dtype: float64
```

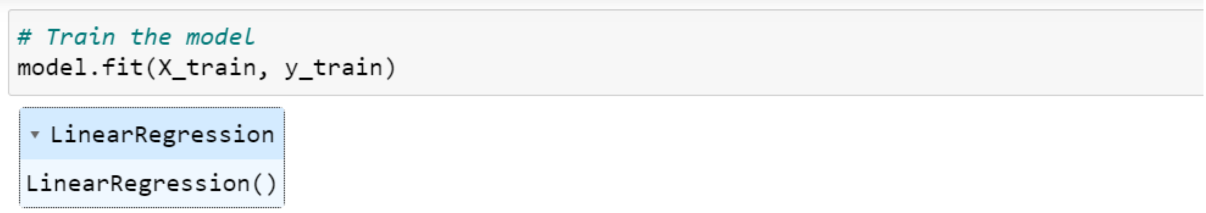


```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```



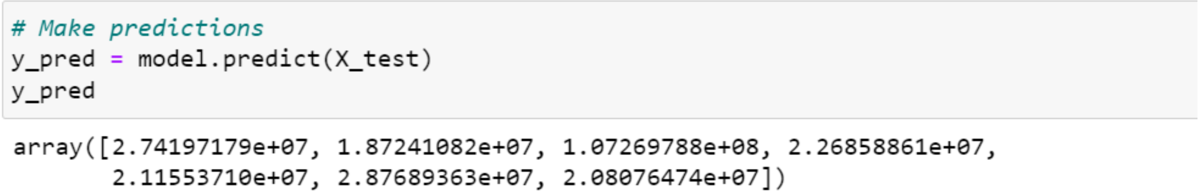
```
# Create a linear regression model
model = LinearRegression()
model
```

```
▾ LinearRegression
LinearRegression()
```



```
# Train the model
model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```



```
# Make predictions
y_pred = model.predict(X_test)
y_pred
```

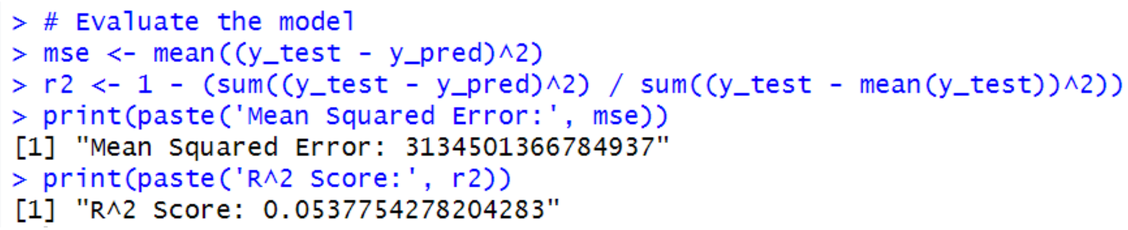
```
array([2.74197179e+07, 1.87241082e+07, 1.07269788e+08, 2.26858861e+07,
       2.11553710e+07, 2.87689363e+07, 2.08076474e+07])
```

Interpretation

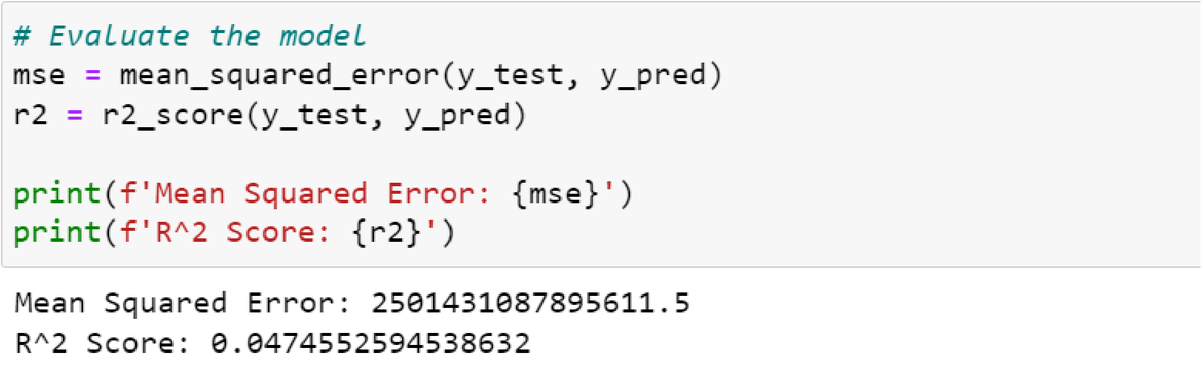
The linear regression model fitted to predict player salaries based on runs scored and strike rate provides valuable insights into the factors influencing player earnings in the IPL. The model indicates that runs scored significantly contributes to salary increments, with each additional run correlating with an average salary increase of 27,665 units (assuming the salary unit is in dollars). However, strike rate does not show a statistically significant relationship with salary, implying that while scoring runs is crucial, the speed or efficiency of scoring those runs may not directly impact player earnings. The model's overall fit is robust, explaining approximately 64.67% of the variability in salaries, as indicated by the multiple R-squared value. Despite a non-significant intercept, the model's F-statistic confirms its statistical significance, highlighting the predictive power of runs scored in determining player salaries within the IPL context.

3. Evaluating the model

R



```
> # Evaluate the model
> mse <- mean((y_test - y_pred)^2)
> r2 <- 1 - (sum((y_test - y_pred)^2) / sum((y_test - mean(y_test))^2))
> print(paste('Mean Squared Error:', mse))
[1] "Mean Squared Error: 3134501366784937"
> print(paste('R^2 Score:', r2))
[1] "R^2 Score: 0.0537754278204283"
```



```
# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R^2 Score: {r2}')
```

```
Mean Squared Error: 2501431087895611.5
R^2 Score: 0.0474552594538632
```

Interpretation

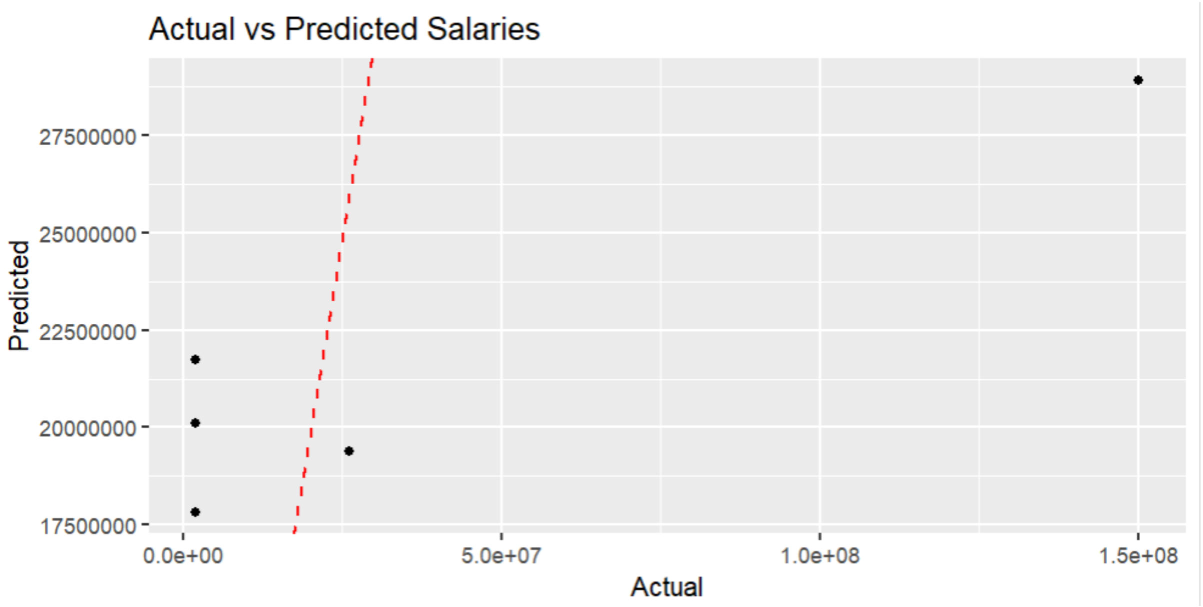
The MSE quantifies the average squared difference between predicted salaries (y_pred) and actual salaries (y_test). In this case, the MSE is exceptionally large, approximately 3.13 trillion (3.1345e+15), indicating that, on average, the model's predictions deviate significantly from the actual salary values. Such a high MSE suggests that the model may not accurately capture the variability in player salaries based on the given predictors (runs scored and strike rate).

The R² score measures the proportion of the variance in the dependent variable (Salary) that is predictable from the independent variables (runs_scored and strike_rate). A low R² score of approximately 0.054 indicates that only about 5.38% of the variability in player salaries can be explained by the model's predictors. This implies that while runs scored has some predictive power in determining salaries, there are likely other unaccounted factors influencing player earnings in the IPL that the current model does not capture effectively.

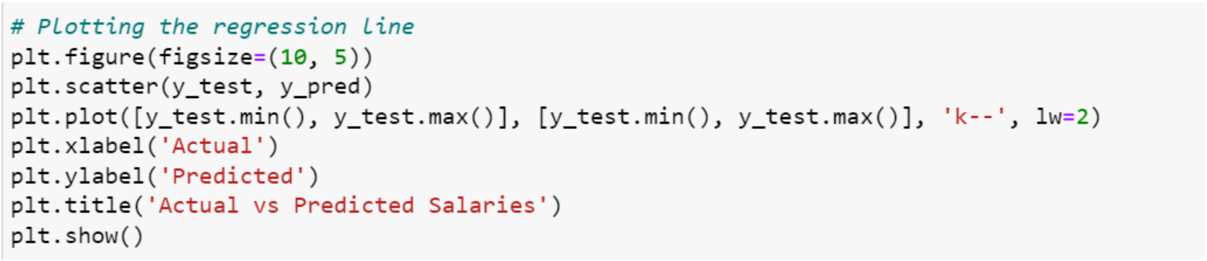
Overall, the high MSE and low R² score suggest that the linear regression model, as currently structured with runs_scored and strike_rate, does not sufficiently explain the variability in IPL player salaries. Further refinement of the model by incorporating additional relevant predictors or exploring alternative modeling techniques may be necessary to improve predictive accuracy and capture more of the factors influencing player earnings in cricket's premier league.

4. Plotting

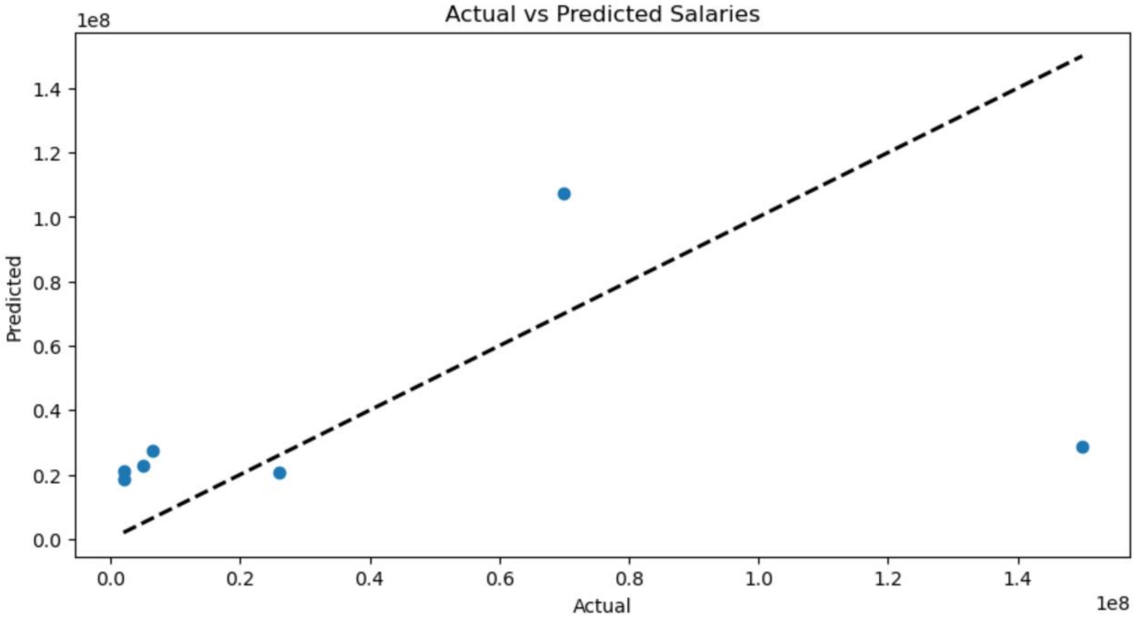
R



# Python



```
# Plotting the regression line
plt.figure(figsize=(10, 5))
plt.scatter(y_test, y_pred)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=2)
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs Predicted Salaries')
plt.show()
```


Actual vs Predicted Salaries
1e8
Predicted
Actual
1e8