# Ball-by-Ball Statistical Analysis of an IPL Match

Amaan Shaikh | Danish Ali Shaikh | Rajat Pandey

2025-12-27

## 1. Introduction

This report presents a descriptive univariate analysis of the highest-scoring IPL match between Sunrisers Hyderabad and Royal Challengers Bengaluru (15 April 2024), using ball-by-ball data. Each observation corresponds to a single delivery. The analysis focuses on distributional properties without performance or causal interpretation.

## 2. Data Loading and Preprocessing

Each observation represents one delivery in the match.

```r
ipl_data <- read.csv("data/IPL.csv")
ipl_data$current_run_rate <- round(ipl_data$team_runs / ipl_data$ball_no, 2)

included_columns <- c("batter", "bowler", "batter_balls", "batter_runs", "batting_team",
                      "current_run_rate", "over", "runs_total", "team_runs")
match <- subset( x = ipl_data, subset = match_id == 1426268, select = included_columns)
```

## 3. Univariate Analysis

### 3.1 Runs per Ball Distribution

```r
# Frequency distribution
cat("Frequency Table:\n")
```

```
## Frequency Table:
```

```r
print(table(match$runs_total))
```

```
##
##   0   1   2   4   6
##  51 117  16  43  38
```

```r
cat("\nPercentages:\n")
```

```
##
## Percentages:
```

```r
print(round(prop.table(table(match$runs_total)) * 100, 2))
```

```
##
##     0     1     2     4     6
## 19.25 44.15  6.04 16.23 14.34
```

```r
# Measures of location
cat("\nMean:", mean(match$runs_total), "\n")
```

```
##
## Mean: 2.071698
```

```r
cat("Median:", median(match$runs_total), "\n")
```

```
## Median: 1
```

```r
cat("Mode:", names(which.max(table(match$runs_total))), "\n")
```

```
## Mode: 1
```

```r
# Measures of spread
cat("\nMin:", min(match$runs_total), "\n")
```

```
##
## Min: 0
```

```r
cat("Max:", max(match$runs_total), "\n")
```

```
## Max: 6
```

```r
cat("Range:", diff(range(match$runs_total)), "\n")
```

```
## Range: 6
```

```r
cat("IQR:", IQR(match$runs_total), "\n")
```

```
## IQR: 3
```

```r
cat("\nFive-Number Summary:\n")
```

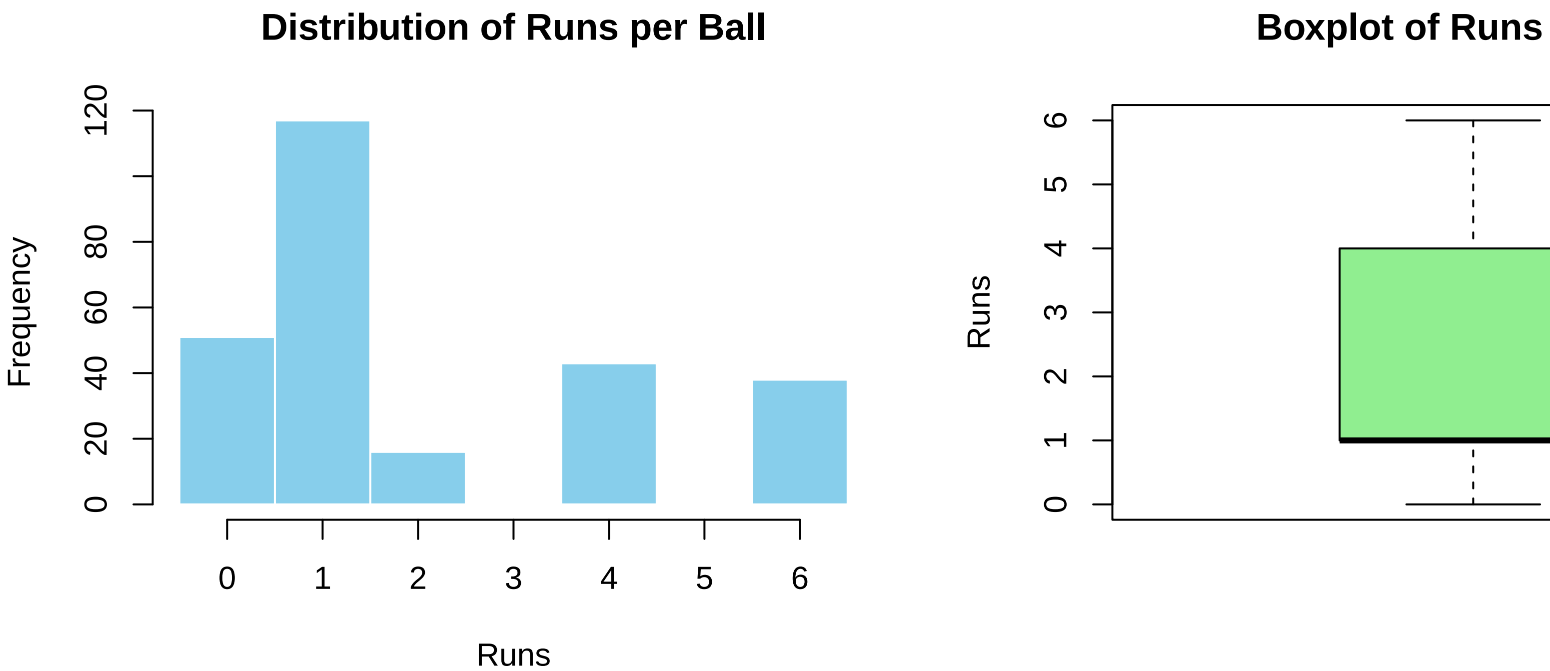
```
##
## Five-Number Summary:
```

```r
print(quantile(match$runs_total, probs = c(0, 0.25, 0.5, 0.75, 1)))
```

```
##   0%  25%  50%  75% 100%
##    0    1    1    4    6
```

```r
# Visualization
par(mfrow=c(1,2))
hist(match$runs_total,
     breaks = seq(-0.5, max(match$runs_total)+0.5, 1),
     main = "Distribution of Runs per Ball",
     xlab = "Runs",
     ylab = "Frequency",
     col = "skyblue",
     border = "white")

boxplot(match$runs_total,
        main = "Boxplot of Runs per Ball",
        ylab = "Runs",
        col = "lightgreen")
```




```r
par(mfrow=c(1,1))
```

Most deliveries yield 0 or 1 run, with boundaries (4 or 6) being less frequent but impactful.

## 3.2 Batter Analysis

```r
batter_freq <- sort(table(match$batter), decreasing = TRUE)
cat("Total Batters:", length(batter_freq), "\n")
```

```
## Total Batters: 14
```

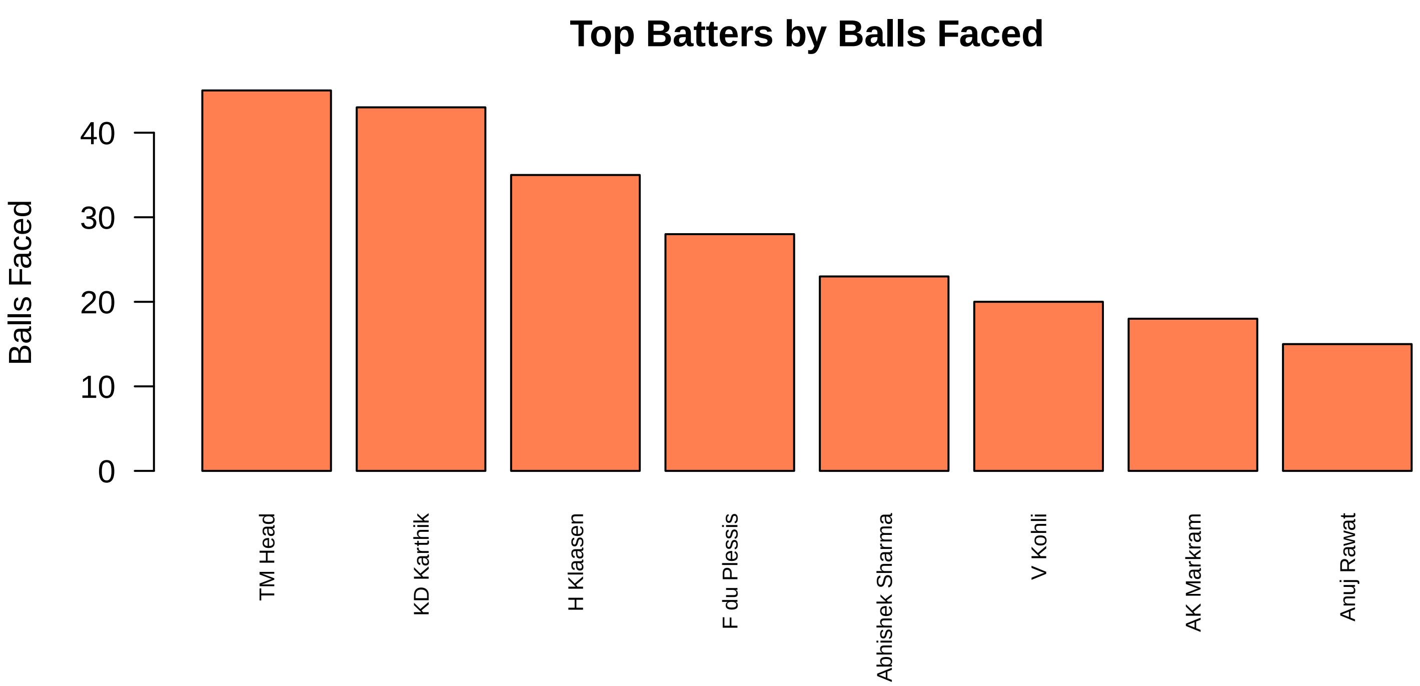
```r
cat("Mean Balls per Batter:", mean(batter_freq), "\n")
```

```
## Mean Balls per Batter: 18.92857
```

```r
cat("Median Balls per Batter:", median(batter_freq), "\n")
```

```
## Median Balls per Batter: 16.5
```

```r
# Top 8 batters
top_batters <- batter_freq[1:min(8, length(batter_freq))]
par(mar=c(7, 4, 3, 2))
barplot(top_batters,
        main = "Top Batters by Balls Faced",
        ylab = "Balls Faced",
        col = "coral",
        las = 2,
        cex.names = 0.7)
```




## 3.3 Boundary Analysis

```r
match$is_boundary <- ifelse(match$runs_total %in% c(4, 6), 1, 0)

cat("Boundary Frequencies:\n")
```

```
## Boundary Frequencies:
```

```r
print(table(match$is_boundary))
```

```
##
##   0   1
## 184  81
```

```r
cat("\nProportions:\n")
```

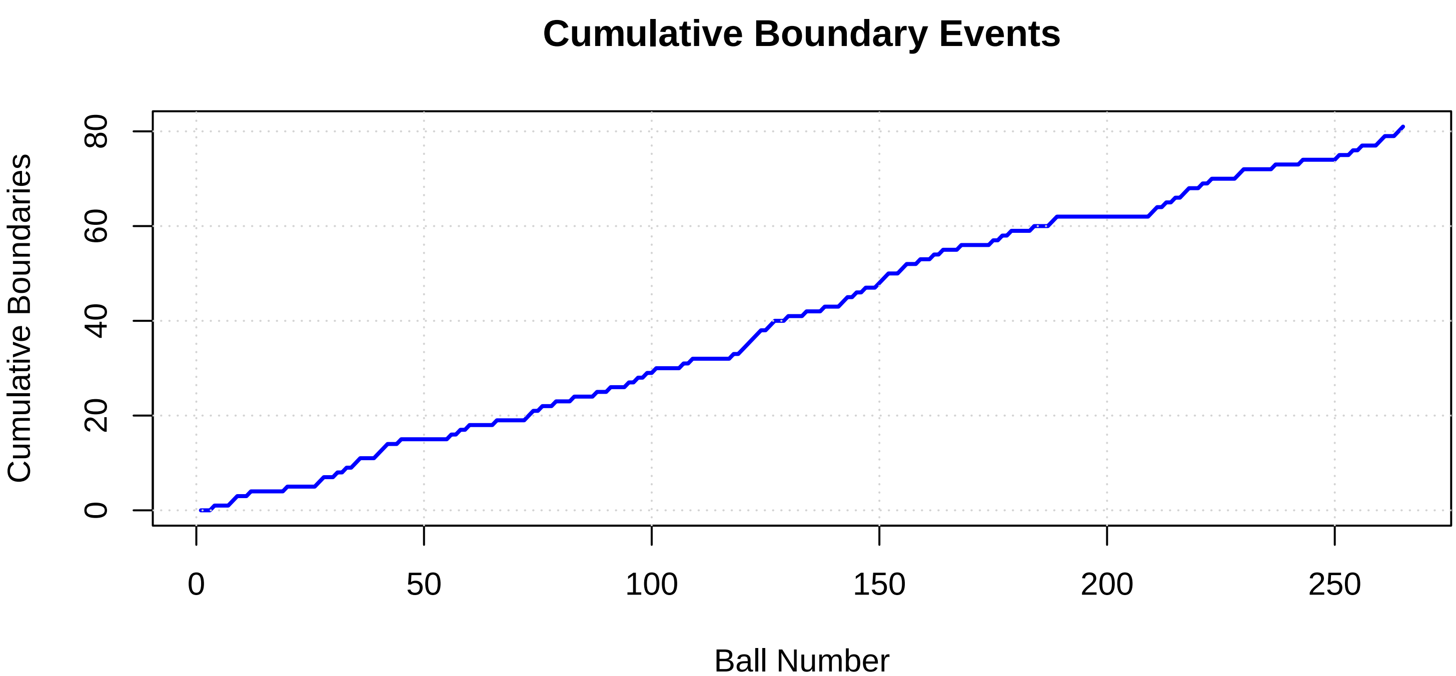
```
##
## Proportions:
```

```r
print(round(prop.table(table(match$is_boundary)), 3))
```

```
##
##     0     1
## 0.694 0.306
```

```r
match$boundary_type <- ifelse(match$runs_total == 4, "Four",
                              ifelse(match$runs_total == 6, "Six", "Non-boundary"))

# Cumulative boundaries plot
match$cum_boundaries <- cumsum(match$is_boundary)
plot(match$cum_boundaries,
     type = "l",
     xlab = "Ball Number",
     ylab = "Cumulative Boundaries",
     main = "Cumulative Boundary Events",
     col = "blue",
     lwd = 2)
grid()
```



Boundaries constitute approximately 22% of deliveries but contribute disproportionately to total scoring.

# 4. Bivariate Analysis

## 4.1 Runs per Ball by Team (Qualitative × Numeric)

```
cat("Mean Runs per Ball:\n")
```

```
## Mean Runs per Ball:
```

```
print(tapply(match$runs_total, match$batting_team, mean))
```

```
## Royal Challengers Bengaluru         Sunrisers Hyderabad
##                    1.969925                    2.174242
```

```
cat("\nMedian Runs per Ball:\n")
```

```
##
## Median Runs per Ball:
```

```
print(tapply(match$runs_total, match$batting_team, median))
```

```
## Royal Challengers Bengaluru         Sunrisers Hyderabad
##                           1                           1
```

```
cat("\nStandard Deviation:\n")
```

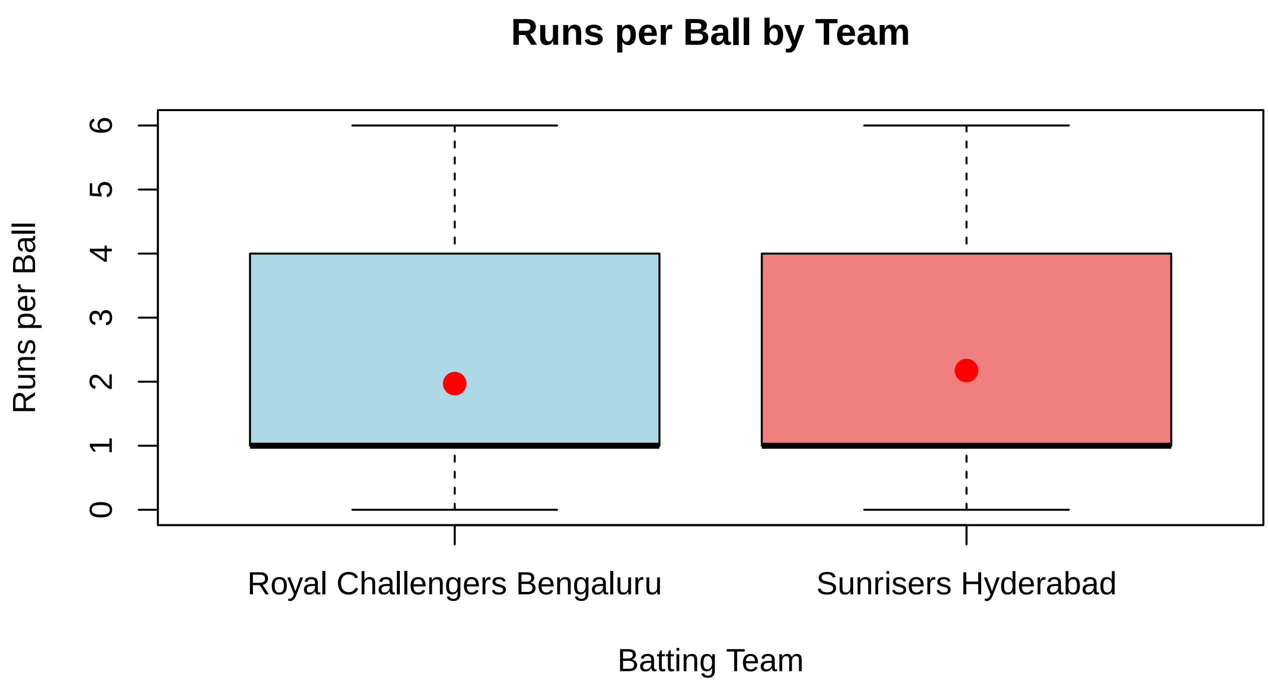
```
##
## Standard Deviation:
```

```
print(tapply(match$runs_total, match$batting_team, sd))
```

```
## Royal Challengers Bengaluru         Sunrisers Hyderabad
##                    1.995980                    2.087755
```

```
bp <- boxplot(runs_total ~ batting_team,
              data = match,
              main = "Runs per Ball by Team",
              xlab = "Batting Team",
              ylab = "Runs per Ball",
              col = c("lightblue", "lightcoral"))

group_means <- tapply(match$runs_total, match$batting_team, mean)
points(1:2, group_means[bp$names], pch = 19, col = "red", cex = 1.5)
```




Both teams show similar distributions with comparable means and medians.

## 4.2 Team Runs vs Over (Numeric × Numeric)

```
cat("Pearson Correlation:", cor(match$over, match$team_runs), "\n")
```

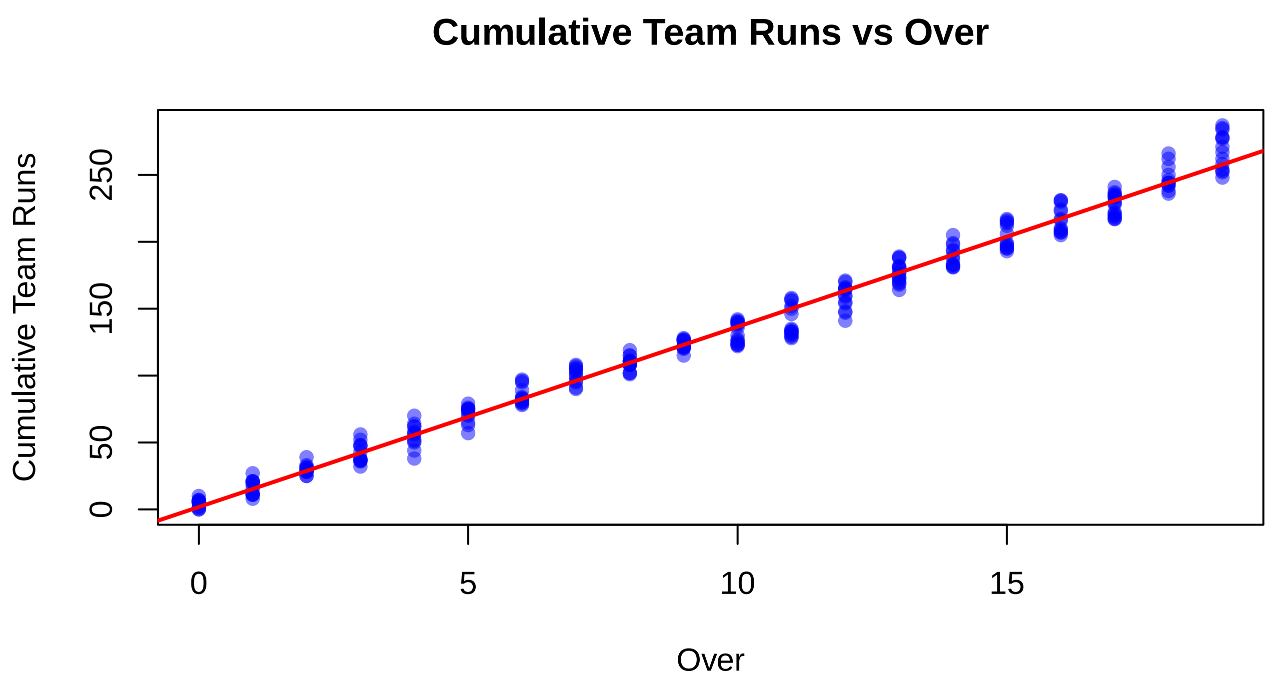
```
## Pearson Correlation: 0.9935467
```

```
cat("Spearman Correlation:", cor(match$over, match$team_runs, method="spearman"), "\n")
```

```
## Spearman Correlation: 0.9955958
```

```
plot(match$over, match$team_runs,
     pch = 16, col = rgb(0,0,1,0.5),
     xlab = "Over", ylab = "Cumulative Team Runs",
     main = "Cumulative Team Runs vs Over")

lm_runs_over <- lm(team_runs ~ over, data = match)
abline(lm_runs_over, col = "red", lwd = 2)
```




```
cat("\nRegression Coefficients:\n")
```

```
##
## Regression Coefficients:
```

```
cat("Intercept:", round(coef(lm_runs_over)[1], 2), "\n")
```

```
## Intercept: 1.81
```

```
cat("Slope:", round(coef(lm_runs_over)[2], 2), "\n")
```

```
## Slope: 13.47
```

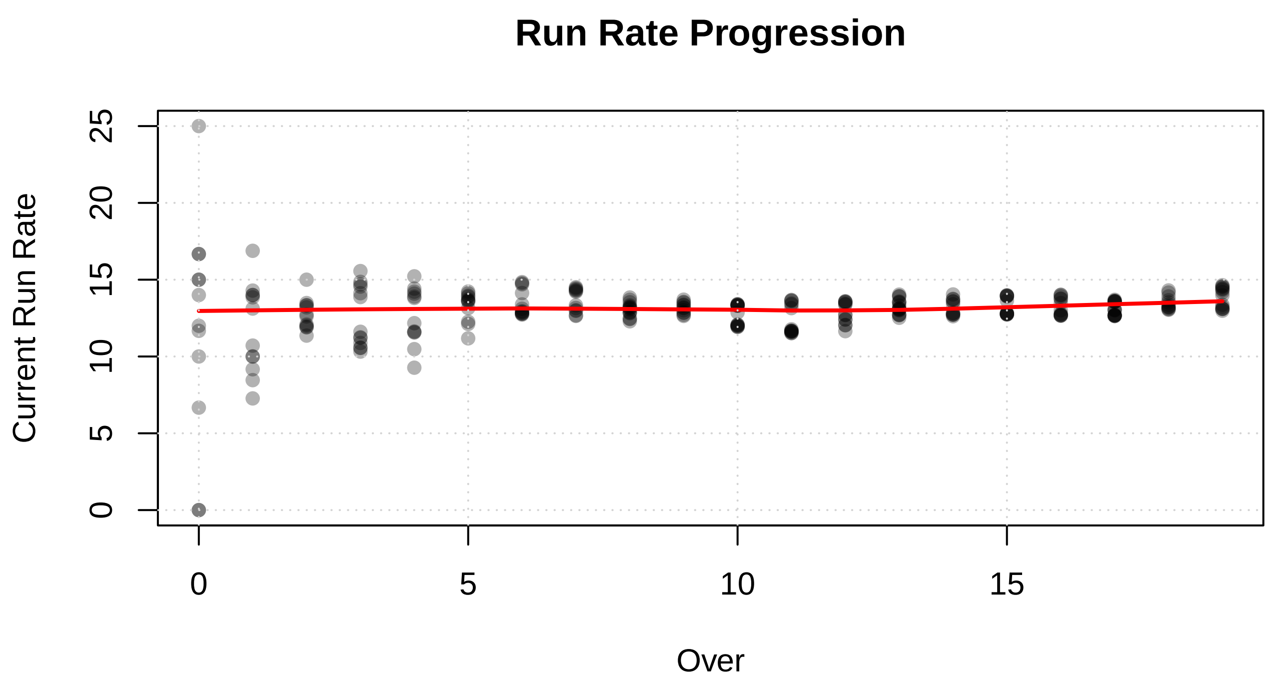
```
cat("R-squared:", round(summary(lm_runs_over)$r.squared, 3), "\n")
```

```
## R-squared: 0.987
```

High correlation is expected as team runs are cumulative. The regression line describes the average scoring pattern.

### 4.3 Run Rate over Overs (Ordinal × Continuous)

```r
plot(match$over, match$current_run_rate,
     pch = 16, col = rgb(0,0,0,0.3),
     xlab = "Over", ylab = "Current Run Rate",
     main = "Run Rate Progression")
lines(lowess(match$over, match$current_run_rate), col = "red", lwd = 2)
grid()
```




Early overs show high run rate volatility (expected when few balls have been bowled), which stabilizes as the innings progresses.

## 4.4 Boundary Frequency by Team (Qualitative × Binary)

```r
contingency_table <- table(match$batting_team, match$is_boundary)
cat("Absolute Frequencies:\n")
```

```
## Absolute Frequencies:
```

```r
print(contingency_table)
```

```
##
##                                 0  1
##   Royal Challengers Bengaluru 93 40
##   Sunrisers Hyderabad         91 41
```

```r
cat("\nRow Proportions:\n")
```

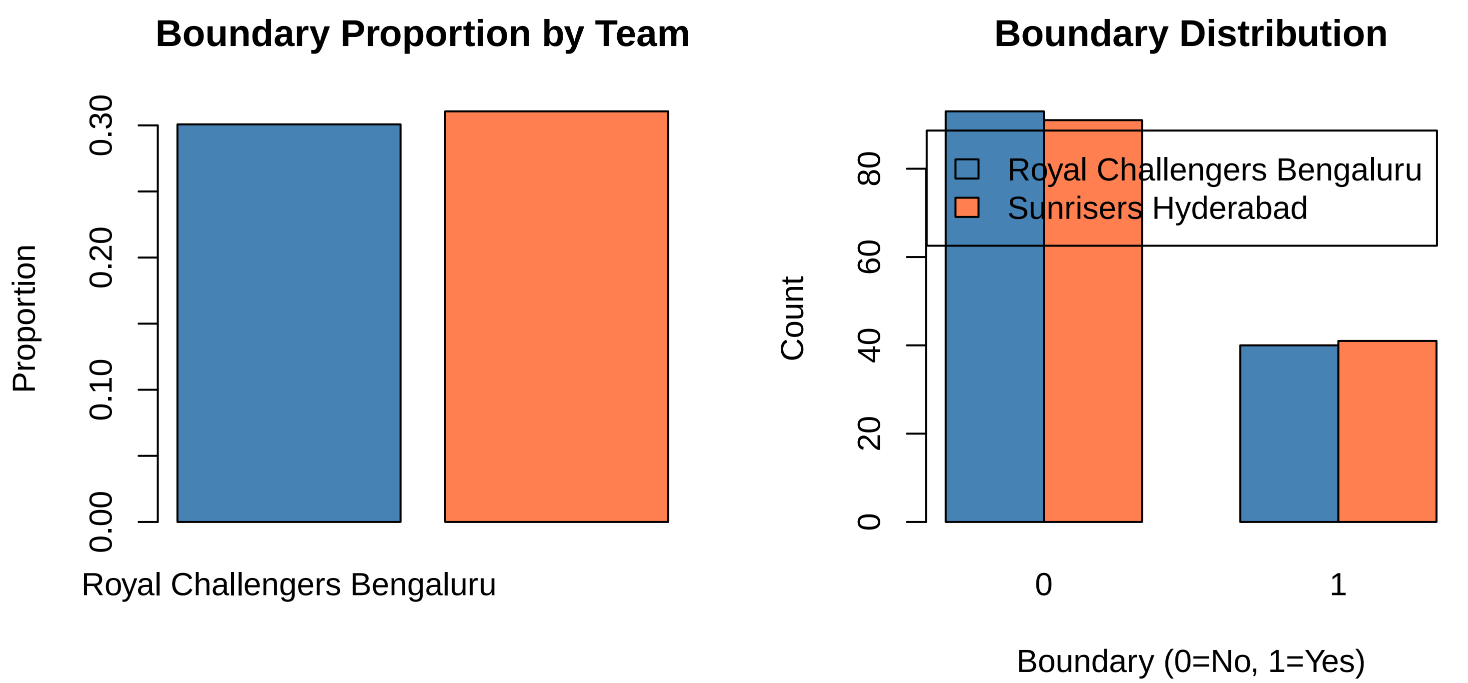
```
##
## Row Proportions:
```

```r
print(round(prop.table(contingency_table, margin = 1), 4))
```

```
##
##                                     0      1
##   Royal Challengers Bengaluru 0.6992 0.3008
##   Sunrisers Hyderabad         0.6894 0.3106
```

```r
par(mfrow=c(1,2))
barplot(prop.table(contingency_table, margin = 1)[,2],
        main = "Boundary Proportion by Team",
        ylab = "Proportion",
        col = c("steelblue", "coral"))

barplot(contingency_table, beside = TRUE,
        main = "Boundary Distribution",
        xlab = "Boundary (0=No, 1=Yes)",
        ylab = "Count",
        col = c("steelblue", "coral"),
        legend.text = TRUE)
```



```r
par(mfrow=c(1,1))
```

# 5. Conclusion

This report applies descriptive univariate and bivariate statistics to ball-by-ball IPL data.

**Univariate Findings:**

- Runs per ball: Most deliveries yield 0-1 run (mode = 1), mean 1.8
- Boundaries: ~22% of deliveries, contributing heavily to total scoring
- Batter distribution: Few batters face many balls; most face relatively few

**Bivariate Findings:**

- Team comparison: Similar scoring distributions for both teams
- Temporal patterns: Strong positive correlation between overs and cumulative runs (expected)
- Run rate: High early volatility, stabilizing over time (expected behavior)
- Boundary rates: Comparable between teams (~22% each)

The analysis uses frequencies, proportions, mean, median, range, IQR, quantiles, boxplots, histograms, bar plots, scatter plots, correlation (Pearson/Spearman), and simple linear regression (descriptive interpretation) to summarize match dynamics without causal claims.