Project Report: Analyzing Team Performance Trends in the IPL: A Statistical and Visual Exploration (2008 – 2023)

---

1. Data Cleaning and Feature Engineering

Summary of Code's Purpose: This section involved loading raw match data and match information, converting dates to a usable format, calculating total runs scored by each team per match, and merging these datasets. Crucially, new features were engineered: a 'Period' column (Pre-2018 vs. Post-2018) for historical comparison, a 'run_rate' column (total runs / 20 overs) for normalized scoring pace, and a 'year' column for season-wise analysis.

Key Output: The **merged** DataFrame, which is the primary dataset for all subsequent analyses, now
contains **match_id**, **batting_team**, **total_runs**, **date**, **venue**, **team1**, **team2**, **Period**, **run_rate**, and **year** columns.

---

2. Shapiro-Wilk Test for Normality

Summary of Code's Purpose: The Shapiro-Wilk test was applied to samples of 'total_runs' from both the 'Pre' (before 2018) and 'Post' (2018 onwards) periods to assess if their distributions follow a normal distribution. This test is crucial for deciding whether to use parametric or non-parametric statistical methods in subsequent analyses.

Key Output:

- **Shapiro-Wilk Test (pre): ShapiroResult(statistic=0.988417764270504, pvalue=0.0005392374071367837)**
- **Shapiro-Wilk Test (post): ShapiroResult(statistic=0.9845150224718922, pvalue=3.6348381261299446e-05)**

Decision on Statistical Methods: For both the 'Pre' and 'Post' periods, the p-values obtained from the Shapiro-Wilk test (0.0005 and 0.000036, respectively) are significantly less than the conventional significance level of 0.05. This leads to the rejection of the null hypothesis that the data is normally distributed. Therefore, the distribution of total runs in IPL matches does **not follow a normal distribution** for either period. Consequently, **non-parametric statistical methods** (such as the Mann-Whitney U test) will be employed for further comparative analyses, as they do not assume normality and are more appropriate for this data.

---

3. & 14. Identification of Playoff Matches

Summary of Code's Purpose: This step involved identifying playoff matches within each IPL season. The methodology assumed that the last four matches of every season, based on their chronological order, constitute the playoffs. The code calculated the total number of matches per season, assigned a sequential match number to each game, and then flagged the final four matches as 'is_playoff = True'. This 'is_playoff' label was then merged into the main **merged** dataset.

Key Output: The **merged** DataFrame now includes a boolean column **is_playoff**, which is **True** for matches identified as playoffs and **False** for regular league matches. This enables direct comparison between league and playoff performances.

---

4. Box Plot: IPL Teams' Total Runs in Playoff vs. League Matches

Summary of Code's Purpose: A box plot was generated to visually compare the distribution of total runs scored by IPL teams in playoff matches versus league matches. This visualization helps in understanding the central tendency, spread, and variability of scores across these two match types.

Key Output: A box plot titled "Team Scores in Playoff vs League Matches" with two boxes, one for 'False' (League) and one for 'True' (Playoff) on the x-axis, showing the distribution of 'total_runs' on the y-axis.

Detailed Analysis of Plot: The box plot indicates that the median total runs scored in playoff matches appears to be slightly lower than in league matches. The interquartile range (the box itself) and the overall spread of scores are also quite similar between the two categories. While there might be a subtle visual difference, it does not suggest a drastic change in scoring patterns in high-stakes playoff games based on this plot alone. The presence of outliers (individual points) indicates some exceptionally high or low scores in both match types.

---

5. Mann-Whitney U Test: Total Runs in Playoff vs. League Matches

Summary of Code's Purpose: A Mann-Whitney U test was intended to be performed to statistically compare the distribution of total runs scored in playoff matches versus league matches. This non-parametric test assesses whether there is a statistically significant difference between the two independent groups.

Key Output:

- **Mann-Whitney U Statistic: 1837444.5**
- **P-value: 1.0**
- **Interpretation: No Statistically significant differnece in scores between playoffs and league matches.**

Detailed Analysis of Output: Upon review of the provided code, there was an error in the data separation for this specific test. Both **league_scores** and **playoff_scores** were inadvertently assigned the same data (league matches). As a result, the Mann-Whitney U test was performed on two identical datasets, which inherently yields a p-value of 1.0. This p-value incorrectly suggests no statistical difference. If the code were corrected to properly separate league and playoff scores, a valid p-value would be generated to determine statistical significance. Based on the erroneous output, no valid conclusion can be drawn from this specific test run.

---

6. Mann-Whitney U Test: Total Runs for a Specific Team (Chennai Super Kings)

Summary of Code's Purpose: This section aimed to apply the Mann-Whitney U test to compare the total runs scored by a specific team, Chennai Super Kings, in their league matches versus their playoff matches. The goal was to determine if CSK's scoring patterns differed significantly in high-stakes games.

Key Output:

- **Team: Chennai Super Kings**
- **Mann-Whitney U Statistic: 19404.5**
- **P-value: 1.0**
- **Interpretation: No Statistically significant differnece in scores between playoffs and league matches.**

Detailed Analysis of Output: Similar to the previous section, the code for this specific team's analysis contained an error where the **playoff_scores** variable was incorrectly assigned the same data as **league_scores**. This led to a p-value of 1.0, which is an invalid result for comparing two distinct groups. Therefore, based on this output, no reliable conclusion can be made regarding the statistical significance of scoring differences for Chennai Super Kings between league and playoff matches.

---

7. Mann-Whitney U Test: Total Runs for Each IPL Team

Summary of Code's Purpose: A Mann-Whitney U test was performed for each unique IPL team to compare their total runs scored in league matches versus playoff matches. The code correctly separated the data for each team into league and playoff categories and ensured sufficient sample sizes before running the test.

Key Output: A list of p-values for each team:

- **Kolkata Knight Riders: p = 0.8206** ❌
- **Royal Challengers Bangalore: p = 0.4337** ❌
- **Chennai Super Kings: p = 0.3327** ❌
- **Kings XI Punjab: p = 0.6591** ❌
- **Delhi Daredevils: p = 0.0892** ❌
- **Rajasthan Royals: p = 0.8971** ❌
- **Mumbai Indians: p = 0.4349** ❌
- **Deccan Chargers: p = 0.0997** ❌
- **Sunrisers Hyderabad: p = 0.7162** ❌
- **Delhi Capitals: p = 0.9859** ❌
- **Gujarat Titans: p = 0.4976** ❌

Detailed Analysis of Output: For all the IPL teams analyzed, the p-values obtained from the Mann-Whitney U test are greater than 0.05. This indicates that for these teams, there is **no statistically significant difference** in their total runs scored between league matches and playoff matches. This suggests that individual teams generally maintain a consistent scoring pattern, and their performance in terms of total runs does not significantly change when moving from regular league games to high-pressure playoff scenarios.

---

8. Violin Plot: Run Rates in Playoff vs. League Matches

Summary of Code's Purpose: A violin plot was created to visualize the distribution of run rates for IPL teams in playoff matches compared to league matches. This plot provides a richer view of the data distribution than a box plot, showing density and spread.

Key Output: A violin plot titled "Run Rate Distribution: Playoffs vs League" showing the distribution of 'run_rate' for 'False' (League) and 'True' (Playoff) on the x-axis.

Detailed Analysis of Plot: The violin plots for both league and playoff matches show very similar distributions of run rates. The central density, median, and overall spread of run rates appear to be consistent across both match types. This visual evidence suggests that the scoring pace of teams does not substantially differ between regular league games and playoff encounters.

---

9. Mann-Whitney U Test: Run Rates in Playoff vs. League Matches

Summary of Code's Purpose: A Mann-Whitney U test was performed to statistically compare the run rates in playoff matches versus league matches. This test evaluates whether any observed difference in scoring rates between the two match types is statistically significant.

Key Output:

- **Mann-Whitney U Statistic: 124606.0**
- **P-value: 0.7668706992141119**
- **Interpretation: ❌ No Statistically significant differnece in scores etween playoffs and league matches.**

Detailed Analysis of Output: The p-value obtained from the Mann-Whitney U test is 0.7669, which is significantly greater than 0.05. This indicates that there is **no statistically significant difference** in the run rates between playoff matches and league matches. This statistical finding supports the visual observation from the violin plot, confirming that the overall scoring pace in IPL matches does not change significantly in playoff scenarios compared to league matches.

---

10. Mann-Whitney U Test: Run Rates for Each IPL Team

Summary of Code's Purpose: A Mann-Whitney U test was conducted for each IPL team to compare their run rates in league matches versus playoff matches. The objective was to

determine if individual teams exhibit statistically significant differences in their scoring pace across these two match types. The results were compiled into a table.

Key Output: A table showing the p-value and significance for each team:

	Team	p-value	Significant
0	Kolkata Knight Riders	0.4976	❌
1	Royal Challengers Bangalore	0.4976	❌
2	Chennai Super Kings	0.4976	❌
3	Kings XI Punjab	0.4976	❌
4	Delhi Daredevils	0.4976	❌
5	Rajasthan Royals	0.4976	❌
6	Mumbai Indians	0.4976	❌
7	Deccan Chargers	0.4976	❌
8	Sunrisers Hyderabad	0.4976	❌
9	Delhi Capitals	0.4976	❌
10	Gujarat Titans	0.4976	❌

Detailed Analysis of Output: For all the teams listed in the output table, the p-values are 0.4976, which is greater than 0.05. This indicates that for every team analyzed, there is **no statistically significant difference** in their run rates between league matches and playoff matches. This suggests that individual teams generally maintain a consistent scoring pace (run rate) regardless of whether they are playing in a regular league game or a high-pressure playoff match.

---

11. Adding Home/Away Columns

Summary of Code's Purpose: Two new boolean columns, **home_team** and **away_team**, were added to the **merged** dataset. These columns indicate whether the **batting_team** in a given match was playing at their designated home ground (**team1**) or as the away team (**team2**). This feature engineering is essential for analyzing the impact of match location on team performance.

Key Output: The **merged** DataFrame now contains **home_team** and **away_team** boolean columns.

---

12. Box Plot: Total Runs Scored by Teams at Home vs. Away

Summary of Code's Purpose: A box plot was generated to visually compare the total runs scored by teams when playing at home versus when playing away. The dataset was first reshaped ('melted') to facilitate this comparison, creating a 'location' column that distinguishes between home and away performances.

Key Output: A box plot titled "Team Scores: Home vs Away" with two boxes, one for 'home_team' and one for 'away_team' on the x-axis, showing the distribution of 'total_runs' on the y-axis.

Detailed Analysis of Plot: The box plot suggests a slight visual difference in total runs scored, with the median for 'home_team' appearing marginally higher than for 'away_team'. This indicates a potential tendency for teams to score more runs when playing at their home ground. The variability and spread of scores are comparable between home and away matches.

---

13. Mann-Whitney U Test: Total Runs Scored at Home vs. Away

Summary of Code's Purpose: A Mann-Whitney U test was performed to statistically assess whether the total runs scored by teams differ significantly when playing at home versus away. This test provides a statistical measure of the performance difference based on match location.

Key Output:

- **Mann-Whitney U Statistic: 581106.0**
- **P-value: 1.2376099361250964e-05** (which is 0.000012376)
- **Interpretation: ❌ The difference in scores between home and away matches is statistically significant.**

Detailed Analysis of Output: The p-value obtained from the Mann-Whitney U test is **1.2376e-05**, which is extremely small and much less than 0.05. This indicates that there is a **statistically significant difference** in the total runs scored by teams when playing at home versus away. This finding suggests that playing at home does provide a statistically significant advantage in terms of scoring more runs.

---

15. Average Total Runs Scored During Different Match Periods (League vs. Playoff)

Summary of Code's Purpose: This section calculates the average total runs scored by each team, separated by 'Period' (Pre/Post 2018), and visualizes these averages using a bar chart. This helps to highlight scoring patterns across teams and different eras of the IPL.

Key Output: A bar plot titled "Top Scoring Teams in Each Period (Average Total Runs per Match)" showing average total runs per team, separated by 'Pre' and 'Post' periods.

Detailed Analysis of Plot: The bar chart displays the average total runs per team, categorized by the 'Pre' and 'Post' 2018 periods. It allows for a visual comparison of how teams' average scores have changed or remained consistent across these two eras. For instance, some teams might show a higher average in the 'Post' period, indicating an overall increase in scoring, while others might remain stable or even decrease. This plot helps identify which teams have adapted their scoring strategies or improved their batting performance over time.

---

16. Match Frequency per Team Across Different Periods (League vs. Playoff)

Summary of Code's Purpose: This analysis calculates how frequently each IPL team appears in matches across the 'Pre' and 'Post' periods (before and after 2018). A bar chart visualizes these match counts, highlighting which teams have consistently participated or played more matches overall in different eras.

Key Output: A bar plot titled "Match Frequency per Team Across Periods" showing the number of matches played by each team, separated by 'Pre' and 'Post' periods.

Detailed Analysis of Plot: The bar chart illustrates the total number of matches played by each team in the 'Pre' and 'Post' periods. This visualization helps to identify long-standing teams that have participated consistently across many seasons versus newer teams or those with shorter tenures. It also provides insight into which teams have played more matches in general, which could correlate with their overall success and longevity in the league.

---

17. Filtering Dataset for Teams with Minimum Matches in Both Periods

Summary of Code's Purpose: This step filters the dataset to include only IPL teams that have played a minimum number of matches (set to 10) in *both* the 'Pre' and 'Post' periods.

This ensures that subsequent comparative analyses are performed on teams with sufficient data in both eras, leading to more meaningful and reliable comparisons.

Key Output:

- **Qualified Teams: ['Chennai Super Kings', 'Delhi Daredevils', 'Kings XI Punjab', 'Kolkata Knight Riders', 'Mumbai Indians', 'Rajasthan Royals', 'Royal Challengers Bangalore', 'Sunrisers Hyderabad']**
- **Filtered Dataset Size: (1669, 13)**

Detailed Analysis of Output: The output clearly lists the 8 teams that met the criteria of playing at least 10 matches in both the 'Pre' and 'Post' periods. The filtered dataset **filtered_merged** now contains 1669 rows and 13 columns, representing only the matches played by these qualified teams. This filtering is crucial for ensuring that any comparisons made between the 'Pre' and 'Post' periods for individual teams are based on a robust amount of data, preventing skewed conclusions from teams with very few matches in one of the periods.

---

18. Average Total Match Score Trend Across IPL Seasons

Summary of Code's Purpose: This analysis calculates the average total match score (combined runs by both teams) for each IPL season from 2008 to 2023. A line plot visualizes this trend over time to identify whether IPL matches have become higher-scoring over the years.

Key Output: A line plot titled "Average Total Match Score by Season" showing 'avg_total_match_score' on the y-axis against 'year' on the x-axis.

Detailed Analysis of Plot: The line plot shows the trend of average total match scores across IPL seasons. It allows for a clear visual assessment of whether there has been an increasing, decreasing, or stable trend in combined team scores over the years. An upward trend would suggest that IPL matches have generally become higher-scoring, potentially due to evolving batting strategies, smaller grounds, or changes in rules.

---

19. Scoring Trends of Major IPL Teams Over the Years

Summary of Code's Purpose: This section calculates each major IPL team's average runs per match per season and visualizes these trends using a line plot. This helps to compare

and analyze how consistent or variable each top team's performance has been across different IPL seasons.

Key Output: A line plot titled "Average Total Match Score by Season" (though the title in the code is "Average Total Match Score by Season", the plot itself shows individual team trends) showing 'avg_score' on the y-axis against 'year' on the x-axis, with different lines for selected top teams (Chennai Super Kings, Mumbai Indians, Royal Challengers Bangalore, Kolkata Knight Riders).

Detailed Analysis of Plot: The line plot displays the average runs scored per match per season for selected major teams. Each line represents a team, showing its scoring trend over the years. This visualization is excellent for comparing the consistency of top teams. For example, some teams might show a steady average, while others might exhibit significant fluctuations, indicating periods of strong or weak batting performance. It helps in understanding the long-term performance trajectory and variability of key franchises.

---

20. Distribution and Relationships of Key IPL Performance Metrics

Summary of Code's Purpose: This analysis examines the statistical properties and relationships of key IPL performance metrics. It calculates the skewness and kurtosis of 'total_runs' and 'run_rate' to understand their distribution shape and presence of outliers. It then converts boolean home/away columns to numeric and generates a correlation matrix to explore relationships between 'total_runs', 'run_rate', 'year', and match location.

Key Output:

- **Skewness and Kurtosis:**
    - **Skewness of total_runs: -0.316126444960667**
    - **Kurtosis of total_runs: 0.597009096235078**
    - **Skewness of run_rate: -0.31612644496066417**
    - **Kurtosis of run_rate: 0.5970090962350754**
- **Correlation Matrix:**

Vorrelation Matrix:

	total_runs	run_rate	year	home_team_int	away_team_int
total_runs	1.000000	1.000000	0.208131	0.096941	-0.096941

run_rate	1.000000	1.000000	0.208131	0.096941	-0.096941
year	0.208131	0.208131	1.000000	0.000238	-0.000238
home_team_int	0.096941	0.096941	0.000238	1.000000	-1.000000
away_team_int	-0.096941	-0.096941	-0.000238	-1.000000	1.000000

Detailed Analysis of Output:

- **Skewness and Kurtosis:**
    - **total_runs** and **run_rate** both show a slight negative skewness (-0.316), indicating that their distributions have a slightly longer tail on the left side (more values concentrated towards the higher end, with fewer very low scores).
    - The kurtosis values (0.597) are positive but less than 3 (for normal distribution), suggesting the distributions are slightly platykurtic (flatter peaks and lighter tails than a normal distribution). This indicates a moderate spread without extreme outliers.
- **Correlation Matrix:**
    - **total_runs** and **run_rate** are perfectly correlated (1.0), which is expected as **run_rate** is directly derived from **total_runs** (total_runs / 20).
    - **year** has a positive correlation with **total_runs** and **run_rate** (0.208), suggesting a weak but positive trend of increasing scores/run rates over the years.
    - **home_team_int** has a small positive correlation with **total_runs** and **run_rate** (0.0969), indicating a slight tendency for higher scores when playing at home. Conversely, **away_team_int** has a negative correlation (-0.0969), as expected.
    - **home_team_int** and **away_team_int** are perfectly negatively correlated (-1.0), which is logical as a team cannot be both home and away simultaneously.

---

21. League-wide Average Runs per Team Per Match by Season

Summary of Code's Purpose: This section calculates the league-wide average runs scored per team per match for each IPL season. A line plot visualizes this trend over time,

helping to identify how scoring patterns have evolved across different seasons at a league level.

Key Output: A line plot titled "League-wise Average Runs per Team per Match by Season" showing 'total_runs' (average) on the y-axis against 'year' on the x-axis.

Detailed Analysis of Plot: The line plot illustrates the overall average runs scored per team per match across all IPL seasons. This provides a macro-level view of scoring trends in the league. An upward trend would suggest that the league as a whole has become higher-scoring over time, potentially due to rule changes, improved batting techniques, or smaller boundaries. Fluctuations might indicate seasons with particularly high or low scoring averages.

---

22. Histogram and KDE of Total Runs per Match by Period

Summary of Code's Purpose: A histogram with overlaid Kernel Density Estimate (KDE) curves was created to compare the distribution of total runs per match between the 'Pre' and 'Post' periods. This visualization helps to visually highlight how scoring patterns differ across these two distinct eras of the IPL.

Key Output: A histogram with KDE curves titled "Histogram and KDE of Total Runs per match by Period" showing the distribution of 'total_runs' for 'Pre' and 'Post' periods.

Detailed Analysis of Plot: The histogram with KDE curves allows for a direct visual comparison of the distribution of total runs in the 'Pre' and 'Post' periods. The 'Post' period distribution appears to be shifted slightly towards higher total runs compared to the 'Pre' period, and its peak might be slightly higher, indicating that matches in the later years of the IPL tend to have higher scores. The KDE curves provide a smoothed representation of these distributions, making it easier to observe the shift and changes in density.

---

23. Hypothesis Testing: Total Runs in Two Periods (League vs. Playoff)

Summary of Code's Purpose: This section performs hypothesis testing to compare total runs between the 'Pre' and 'Post' periods. It uses the Mann-Whitney U test to check for statistically significant differences in distributions and calculates Cohen's d to measure the effect size of the difference.

Key Output:

- **Mann-Whitney U Test: U-statistic = 381621.5 , p-value = 1.454889995517028e-17**

- **Cohen's d effect Size: 0.014118121481058251**

Detailed Analysis of Output:

- **Mann-Whitney U Test:** The p-value is **1.45e-17**, which is extremely small (virtually zero) and much less than 0.05. This indicates a **highly statistically significant difference** in the distribution of total runs between the 'Pre' and 'Post' periods. This confirms that scoring patterns have indeed changed significantly over time in the IPL.
- **Cohen's d Effect Size:** Cohen's d is **0.0141**. This value represents a **very small effect size**. While the difference is statistically significant (due to a large sample size), the practical significance of this difference in total runs between the 'Pre' and 'Post' periods is minimal. This means that although the distributions are statistically different, the magnitude of that difference in average runs is not large enough to be considered practically meaningful in a real-world context.

---

24. Bootstrapping for Confidence Interval of Mean Difference

Summary of Code's Purpose: Bootstrapping was used to estimate a 95% confidence interval for the difference in average total runs between the 'Pre' and 'Post' periods. This non-parametric method repeatedly resamples the data to build a distribution of the mean difference, providing a robust range within which the true mean difference likely falls.

Key Output:

- **95% Confidence Interval: [9.33439205 14.51650406]**
- **Mean Difference: 11.955175179070357**

Detailed Analysis of Output: The 95% confidence interval for the difference in average total runs between the 'Pre' and 'Post' periods is approximately **[9.33, 14.52]**. The mean difference is about **11.96** runs. This means we are 95% confident that the true average total runs in the 'Post' period are between 9.33 and 14.52 runs higher than in the 'Pre' period. This provides a more intuitive understanding of the magnitude of the difference, complementing the Cohen's d result by showing the actual range of the difference in runs.

---

25. Linear Regression Model: Trend in Average Total Runs per Match Over Seasons

Summary of Code's Purpose: A linear regression model was fitted to analyze the trend in average total runs per match over IPL seasons. 'Year' was used as the independent variable

and 'average total runs' as the dependent variable. The statistical summary of the model helps assess whether scoring has increased or decreased over time and the strength of that trend.

Key Output:

RunCopy code

OLS Regression Results

==========

Dep. Variable: total_runs R-squared: 0.647

Model: OLS Adj. R-squared: 0.622

Method: Least Squares F-statistic: 25.68

Date: Fri, 18 Jul 2025 Prob (F-statistic): 0.000172

Time: 11:06:04 Log-Likelihood: -47.087

No. Observations: 16 AIC: 98.17

Df Residuals: 14 BIC: 99.72

Df Model: 1

Covariance Type: nonrobust

==========

coef std err t P>|t| [0.025 0.975]

---

const -2569.0983 536.414 -4.789 0.000 -3719.592 -1418.605

year 1.3488 0.266 5.068 0.000 0.778 1.920

==========

Detailed Analysis of Output:

- **R-squared (0.647):** Approximately 64.7% of the variance in average total runs can be explained by the 'year' (season). This indicates a moderately strong linear relationship.
- **P-value for 'year' (0.000):** The p-value for the 'year' coefficient is 0.000, which is highly statistically significant (much less than 0.05). This means there is a significant linear trend in average total runs over the seasons.
- **Coefficient for 'year' (1.3488):** The positive coefficient of 1.3488 indicates that, on average, the total runs per match have increased by approximately **1.35 runs per season**.
- **Conclusion:** The linear regression model suggests a **statistically significant upward trend** in average total runs per match over the IPL seasons. IPL matches have generally become higher-scoring over the years, and this trend is moderately strong.

---

26. Linear Regression Model: Prediction and Visualization of Trend

Summary of Code's Purpose: This section uses the fitted linear regression model to predict average IPL team scores over the years. It then visualizes both the actual seasonal averages (scatterplot) and the predicted linear trend (red line) on a single plot. This helps to visually assess how team scoring has changed over time and how well the linear model captures this trend.

Key Output: A scatter plot showing actual average runs per season, with an overlaid red line representing the fitted linear trend. The plot is titled "Average Runs Per Match by Team (Over Years)".

Detailed Analysis of Plot: The scatter plot shows the actual average runs per match for each season, while the red line represents the linear trend predicted by the regression model. The red line clearly shows an upward slope, visually confirming the increasing trend in average runs per match over the years. The scatter points generally cluster around this line, indicating that the linear model provides a reasonable fit for the observed trend, although there are some deviations in individual seasons. This visualization effectively communicates that IPL matches have indeed become higher-scoring over the period analyzed.
