PSL Analytics: Deciphering 9 Years of Cricket Match Data

ICT (CS202)
Project Report

Submitted by:

Raja Hamza Sikandar	2024532
Section:	H (BCS)

Submitted to:

M Talha Ashfaq

Unveiling Victory Patterns: A Statistical Analysis of PSL Trends (2016–2024) using R

I. Dataset Description

For this project, I chose a relatable dataset: **Pakistan Super League (PSL) 2016-2024** [1]. This dataset contains the comprehensive records of all matches played in the HBL PSL from 2016 to 2024.

A. Features and Structure

The dataset was provided as a CSV file and contains a mix of features:

- **Numerical Features:** `Win By Runs`, `Win By Wickets`, etc.
- **Categorical Features:** `Team 1`, `Team 2`, `Toss Decision`, etc.
- **Target Variable:** I chose `Winner` as my target variable. This allows for modeling the effect of various factors (like Toss or Venue) on the match outcome.

II. Data Importing and Preprocessing

The raw dataset required several cleaning steps to ensure quality analysis:

- **Handling Missing Values:** A significant number of missing values were found in the `Win by Runs` and `Win by Wickets` columns. This is inherent to cricket logic: if a team wins by wickets, the ”runs” margin is typically empty. Instead of removing these rows, I imputed 0 for the missing values, preserving valid match data.
- **Duplicate Removal:** I used the `duplicated()` function to identify and remove duplicate rows to prevent incorrect statistics.
- **Categorical Encoding:** Features such as `Team 1`, `Toss Decision`, and `Winner` were converted from character strings to factors to help in statistical plotting and future modeling.

III. Key Findings from EDA

A. Summary Statistics

Analyzing the numerical features showed distinct patterns in victory margins. The Standard Deviation for `Win By Runs` was calculated to summarize the variability of defending targets. On the other hand, the stats for `Win By Wickets` showed how comfortable teams generally are while chasing (likely due to knowing the required score).

B. Feature Distributions

Target Variable (Winner): The bar plot of the Winner column revealed the dominance hierarchy in the PSL. Teams like **Islamabad United** and **Peshawar Zalmi** appeared frequently as winners, while newer teams had different win counts due to playing fewer seasons.

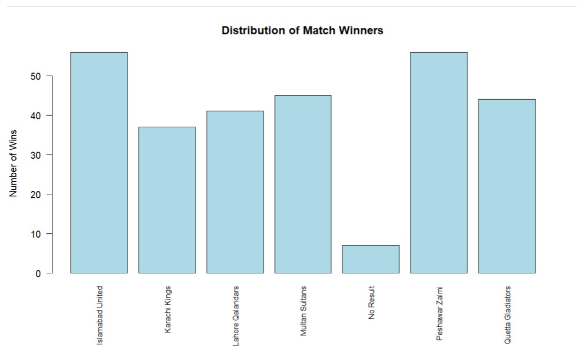

Fig. 1. Distribution of Match Winners. Islamabad United and Peshawar Zalmi show consistent performance.

Toss Decision: The frequency distribution highlighted a strategic preference. Teams overwhelmingly chose to **Field First (Bowl)** rather than Bat First. This reflects the modern T20 trend of preferring to chase targets and the nature of Pakistani pitches, where batting usually gets easier in the second innings.

IV. Insights from Visualizations

A. Correlation Matrix

The correlation matrix highlighted a trivial but important negative relationship between `Win by Runs` and `Win by Wickets`. Since a match can only be won by one method, these two variables are mutually exclusive.

More interestingly, the correlation between `Toss Winner` and `Winner` (visualized via the new feature `Toss_Win_Match_Win`) provided insight into how often the ”luck of the toss” translates to a match victory.

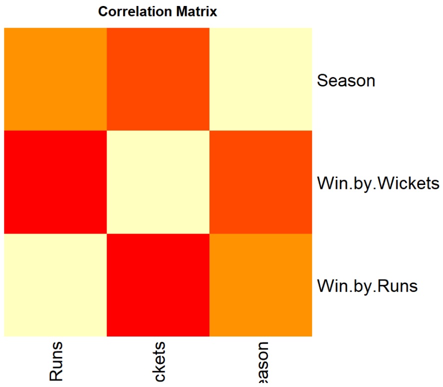

Fig. 2. Correlation Matrix. The heatmap confirms the mutual exclusivity of run and wicket margins.

B. Scatter Plots

The scatter plot of `Win by Runs` vs. `Win by Wickets` formed a distinct ”L-shape” along the axes, confirming the mutual exclusivity of the victory types. Additionally, the `Season` vs. `Win by Runs` visualization helped track scoring trends over the years.

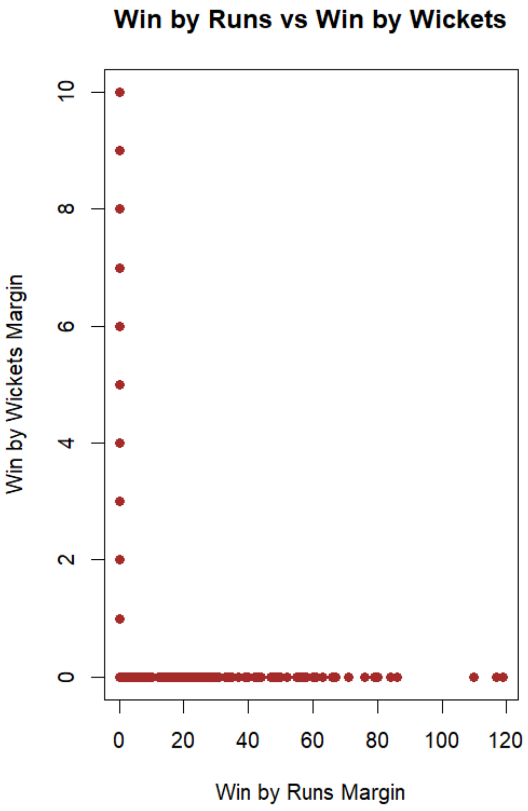

Fig. 3. Scatter Plot of Win by Runs vs Wickets, showing the distinct ”L-shape” distribution.

C. Boxplots

Boxplots comparing `Win by Runs` across different `Winner` classes revealed team-specific characteristics. For instance, some teams showed a higher median run margin, suggesting they are stronger at defending totals compared to others.

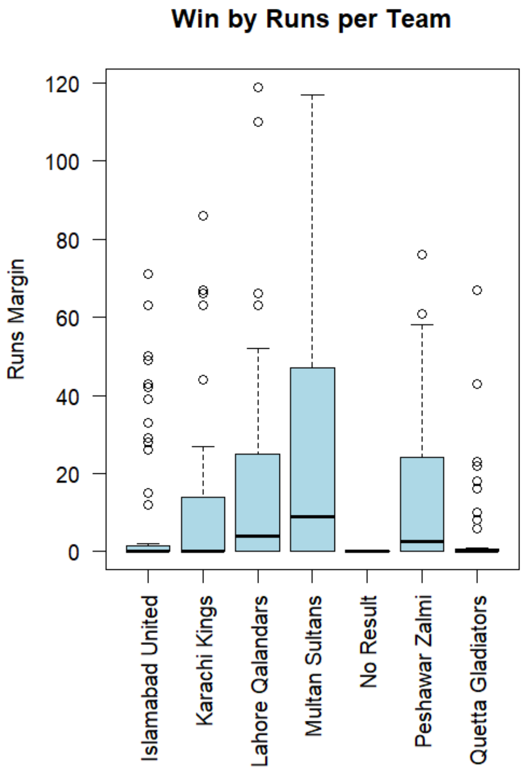

Fig. 4. Boxplot of Win by Runs per Team. Note the varying medians for different franchises.

V. Feature Engineering

To improve potential model performance, I applied the following transformations:

- **Scaling:** The `Win by Runs` (scale 0-100) and `Win by Wickets` (scale 0-10) features were standardized using Z-score scaling (`scale()`). This ensures that a distance-based algorithm (like k-NN) would not be biased toward run margins.
- **New Features:**
  - `Win_Type`: Categorized victories into ”Defending” (Runs) or ”Chasing” (Wickets).
  - `Toss_Win_Match_Win`: A binary feature indicating if the toss winner also won the match, which is a strong predictor in T20 cricket analysis.

VI. Next Steps

Based on this EDA, the following steps are recommended for further analysis:

- **Predictive Modeling:** Train a Random Forest or Logistic Regression model using the training set (70% split) to predict the Winner.

- **Hypothesis Testing:** Perform a T-test to statistically confirm if teams batting second have a significantly higher win rate than teams batting first.
- **Advanced Feature Creation:** Integrate external data such as ”Player of the Match” impact or ”Venue Scoring Rates” to improve prediction accuracy.

References

[1] Kaggle, ”Pakistan Super League (PSL) 2016-2024 Dataset,”.