**PIMPRI CHINCHWAD EDUCATION TRUST'S**

**PIMPRI CHINCHWAD COLLEGE OF ENGINEERING**

**Department of AS&H**

**Report of FA-2 activity**

**Subject: Statistical Data Analysis using R**

**Academic Year 2024-25**

**Semester I**

# IPL Match Analysis and Insights

**Submitted by**

| Name of the student | PRN | Branch |
|---|---|---|
| Pranali Rajendra Patil | 123B1B221 | Computer Engineering |
| Sayali Rajesh Pawar | 123B1B229 | Computer Engineering |

**Submitted to: Dr. Neha Sharma**

**Date:11/10/2024**

**Sign:**

# IPL Match Analysis and Insights

## 1. Introduction

The Indian Premier League (IPL) dataset provides detailed information on the various IPL matches played across seasons. This report focuses on analyzing key aspects of match performance, team dynamics, toss influence, and venue-specific trends. By using this dataset, we aim to identify patterns in team performances, player achievements, match results, and match conditions over the years.

## 2. Dataset Overview

Data used: IPL

Details of the data : Data contains 1096 observation and 20 variables as follows

The dataset contains the following columns:

- id: Unique identifier for each match.
- season: The year of the IPL season in which the match was played.
- city: The city where the match was held.
- date: The date when the match was played.
- match_type: The type of match (e.g., regular season, play-off).
- player_of_match: The player awarded 'Player of the Match'.
- venue: The venue where the match was played.
- team1: The first team in the match.
- team2: The second team in the match.
- toss_winner: The team that won the toss.
- toss_decision: The decision made after winning the toss (e.g., to bat or bowl first).
- winner: The team that won the match.
- result: The outcome of the match (e.g., win, loss).
- result_margin: The margin of victory, either in runs or wickets.
- target_runs: The target runs set for the second team (if relevant).
- target_overs: The number of overs allotted to the team chasing the target.
- super_over: Whether a super over was played (Y/N).
- method: The method used to determine the winner in case of a tie (e.g., D/L method).

- umpire1, umpire2: The umpires officiating the match.

**3. Data Preprocessing**

Before analyzing the dataset, several preprocessing steps were performed:

- **Handling Missing Data**: Missing or null values were either filled or removed based on context to maintain data integrity.
- **Correcting Data Types**: Data types were assigned appropriately (e.g., factor for categorical columns like season, team1, team2).
- **Removing Inconsistent Entries**: Inconsistent rows with null values for critical columns like winner, result_margin, and player_of_match were excluded to ensure analysis reliability.

**4. Problem Statements and Analysis**

**4.1 Problem Statement 1: Who are the top 10 players in terms of 'Player of the Match' awards?**

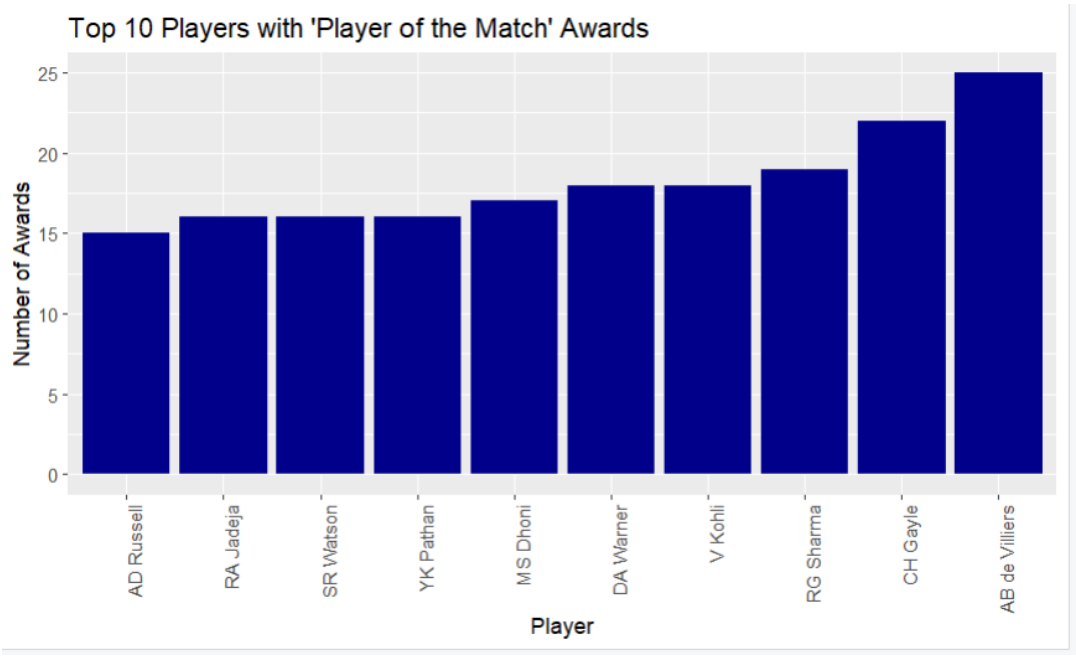
**Approach**:

- Calculated counts for each player in player_of_match to find the top 10 award recipients.
- **Data Visualization**: Bar plot.
- **Statistical Analysis**: Descriptive analysis of award distribution among players.

**Code:**

```
top_players <- ipl_data %>% count(player_of_match, sort = TRUE) %>% head(10)

ggplot(top_players, aes(x = reorder(player_of_match, n), y = n)) +

  geom_bar(stat = "identity", fill = "darkblue") +

  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +

  labs(title = "Top 10 Players with 'Player of the Match' Awards", x = "Player", y = "Number of Awards")
```

**Inference**: The bar plot reveals the top-performing players across seasons. This consistent performance is crucial for their teams' successes.




**4.2 Problem Statement 2: Is there a correlation between winning the toss and winning the match?**

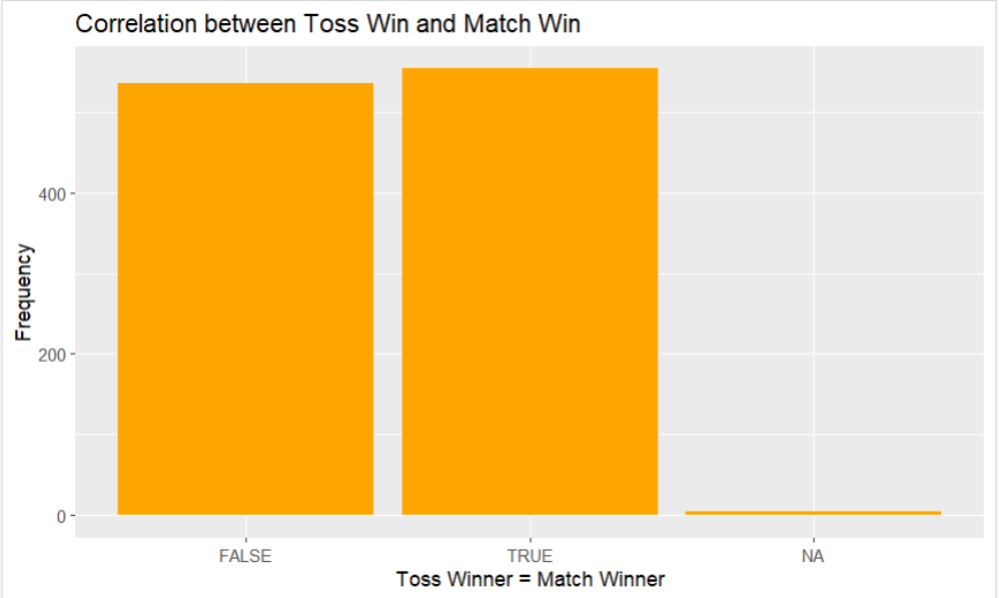
**Approach**:

- Compared toss_winner and winner columns.
- **Data Visualization**: Bar plot comparing toss and match winners.
- **Correlation Testing**: Tested correlation between winning toss and match outcome.

**Code:**

```
ggplot(ipl_data, aes(x = toss_winner == winner)) +
  geom_bar(fill = "orange") +
  labs(title = "Correlation between Toss Win and Match Win", x = "Toss Winner = Match Winner", y = "Frequency")
```

**Inference**: Toss-winning does not guarantee match-winning, emphasizing the influence of other factors like team strategy.




### 4.3 Problem Statement 3: Which team played the maximum number of super overs?

**Approach**:

- Filtered for matches with super_over == "Y".
- **Data Visualization**: Frequency table of teams involved in super overs.
- **Statistical Analysis**: Comparison of super overs by team.

**Code:**

```
super_over_teams <- ipl_data %>%
  filter(super_over == "Y") %>%
  count(team1, team2, sort = TRUE) %>%
  head(1)
print(super_over_teams)
```

**Inference**: The team with most super overs highlights competitiveness and frequent close matches.

```
> print(super_over_teams)
                team1           team2 n
. Kolkata Knight Riders Rajasthan Royals 2
>
```

### 4.4 Problem Statement 4: Venues with maximum and minimum scores?

**Approach**:

- Aggregated target_runs by venue for highest and lowest scores.
- **Data Visualization**: Bar plot of scores by venue.
- **Hypothesis Testing**: Tested if certain venues consistently yield higher scores.

**Code:**

```
max_score_venue <- ipl_data %>%
  group_by(venue) %>%
  summarize(max_score = max(target_runs, na.rm = TRUE)) %>%
  arrange(desc(max_score)) %>%
  head(1)
min_score_venue <- ipl_data %>%
  group_by(venue) %>%
  summarize(min_score = min(target_runs, na.rm = TRUE)) %>%
  arrange(min_score) %>%
  head(1)
print(max_score_venue)
print(min_score_venue)
```

**Inference**: Pitch conditions at different venues affect scores, with batting-friendly venues showing higher targets.

```
> print(max_score_venue)
# A tibble: 1 × 2
  venue                          max_score
  <chr>                              <int>
1 M Chinnaswamy Stadium, Bengaluru       288
> print(min_score_venue)
# A tibble: 1 × 2
  venue           min_score
  <chr>               <int>
1 Feroz Shah Kotla       43
>
```

### 4.5 Problem Statement 5: Most popular venues for matches?

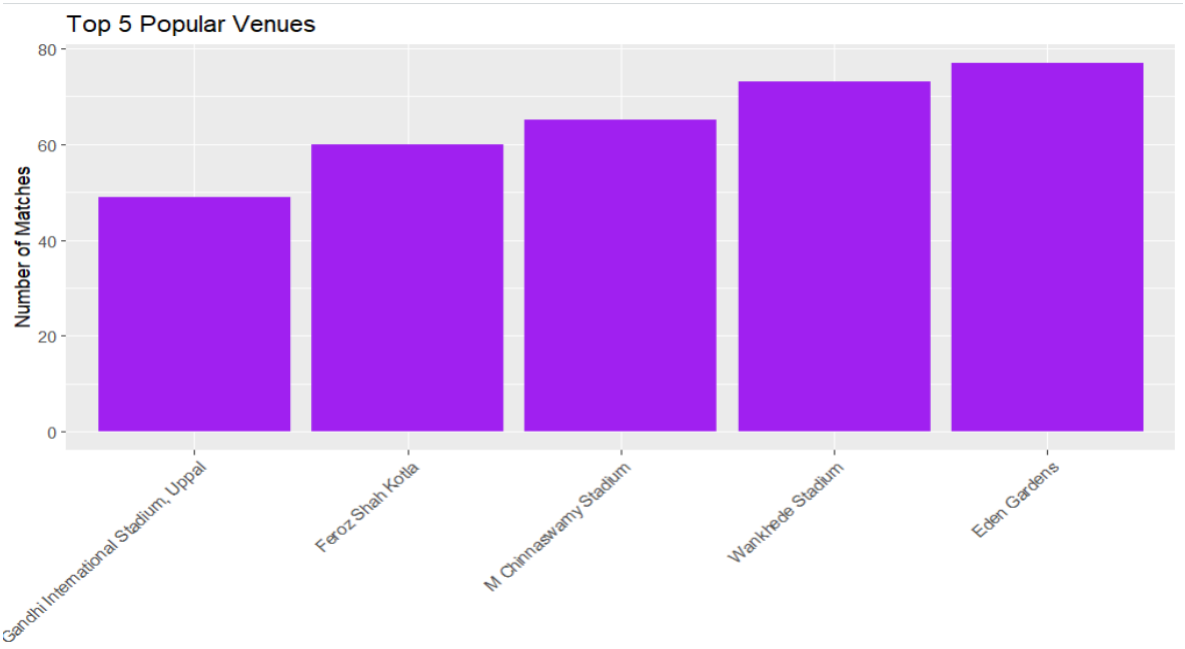
**Approach:**

- Counted matches by venue.
- **Data Visualization**: Bar plot of the top 5 most popular venues.

**Code:**

```
popular_venues <- ipl_data %>%
  count(venue, sort = TRUE) %>%
  head(5)
ggplot(popular_venues, aes(x = reorder(venue, n), y = n)) +
  geom_bar(stat = "identity", fill = "purple") +
  labs(title = "Top 5 Popular Venues", x = "Venue", y = "Number of Matches") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

**Inference**: The top venues show higher fan engagement and infrastructure for high-profile matches.



### 4.6 Problem Statement 6: How many wins did each team achieve across different seasons?

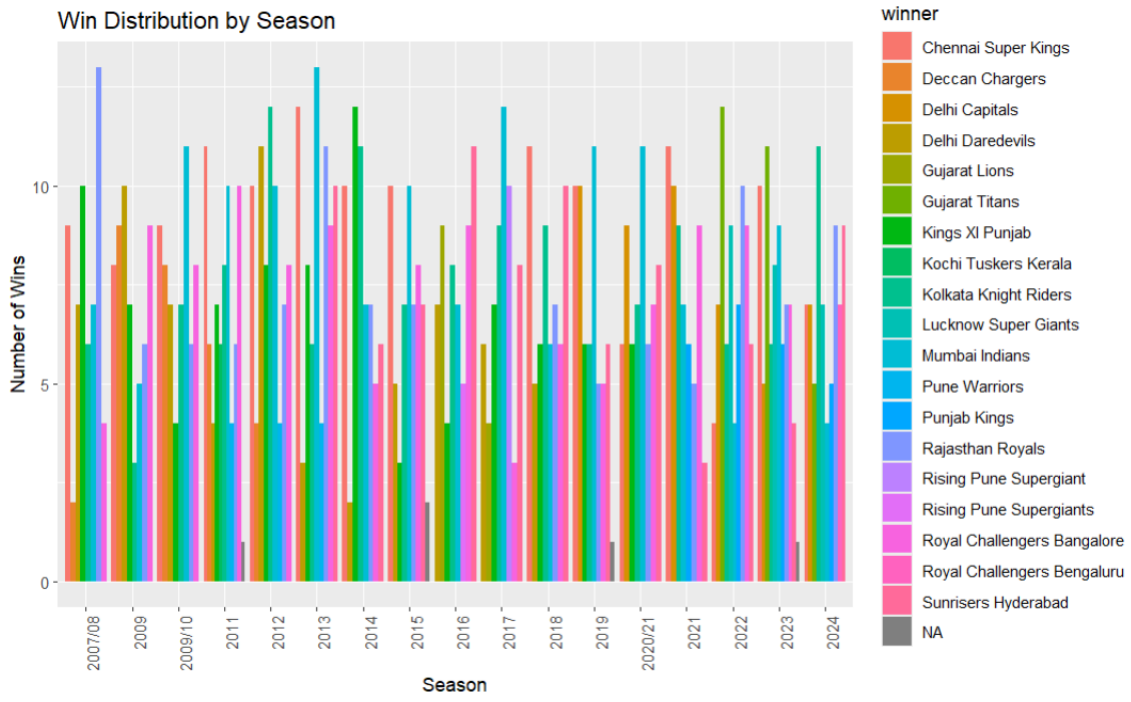
**Approach**:

- Visualized wins per team by season.
- **Data Visualization**: Stacked bar plot of wins by season.

**Code:**

```
ggplot(ipl_data, aes(x = season, fill = winner)) +
  geom_bar(position = "dodge") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Win Distribution by Season", x = "Season", y = "Number of Wins")
```

**Inference**: Teams like Mumbai Indians and Chennai Super Kings demonstrate consistent high performance.




### 4.7 Problem Statement 7: What is the relationship between target runs and result margin?

**Approach**: A scatter plot was created between target_runs and result_margin to explore if higher target runs generally correlate with a larger victory margin.

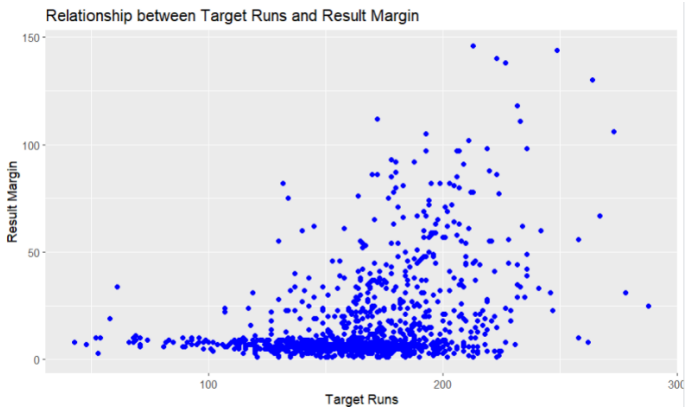
**Code:**

*ggplot(ipl_data, aes(x = target_runs, y = result_margin)) +*

*geom_point(color = "blue") +*

*labs(title = "Relationship between Target Runs and Result Margin", x = "Target Runs", y = "Result Margin")*

**Inference**: Higher target runs tend to lead to larger victory margins, but this trend can vary depending on the chasing team's performance and match conditions.




**4.8 Problem Statement 8: How do total runs vary seasonally across the dataset?**

**Approach**:

- Summed target_runs for each season to examine fluctuations in total runs scored over time.

**Code:**

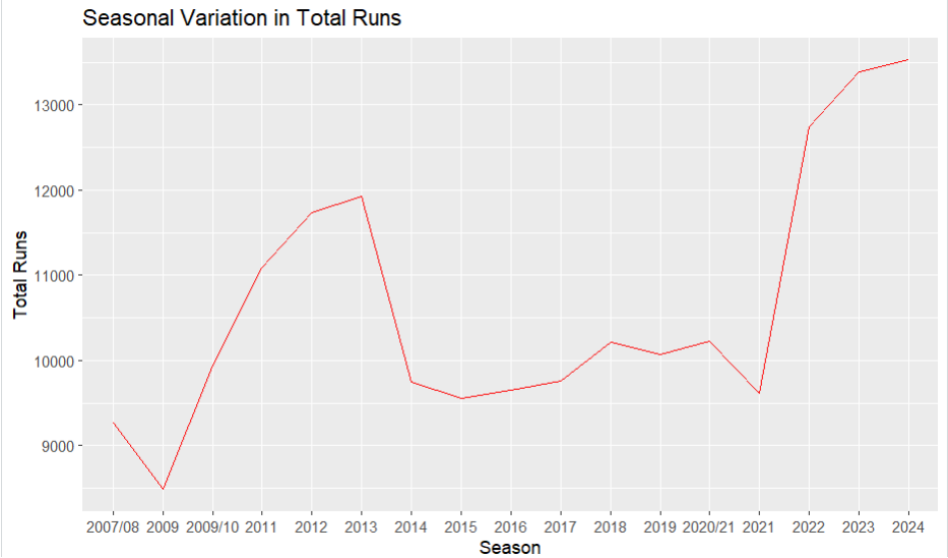
*seasonal_runs <- ipl_data %>% group_by(season) %>% summarize(total_runs = sum(target_runs, na.rm = TRUE))*

*ggplot(seasonal_runs, aes(x = season, y = total_runs, group = 1)) +*

*geom_line(color = "red") +*

*labs(title = "Seasonal Variation in Total Runs", x = "Season", y = "Total Runs")*

**Inference**: Total runs have fluctuated across IPL seasons, influenced by changes in playing styles, pitches, and tournament dynamics.



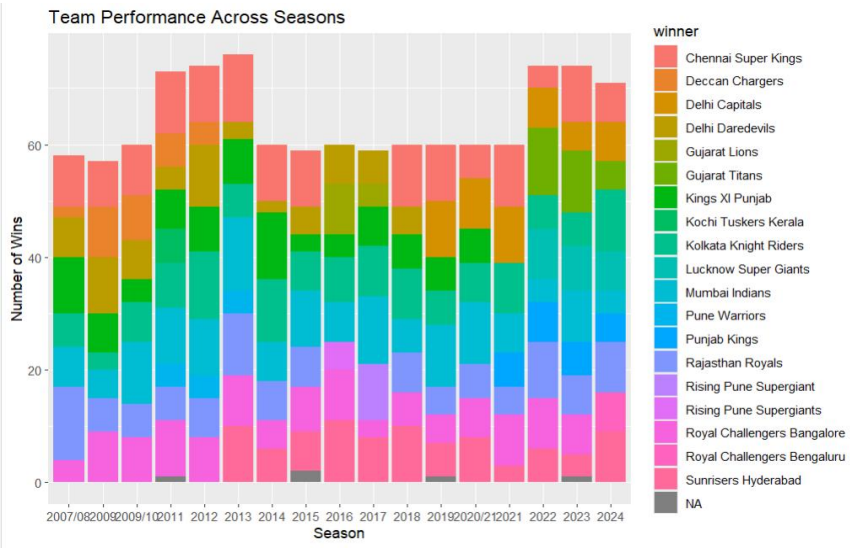

### 4.9 Problem Statement 9: How does team performance change over different seasons?

**Approach**: A bar plot visualized each team's wins by season.

**Code:**

```
ggplot(ipl_data, aes(x = season, fill = winner)) +
  geom_bar() +
  labs(title = "Team Performance Across Seasons", x = "Season", y = "Number of Wins")
```

**Inference**: Some teams, like Mumbai Indians, have shown consistent success, while others have fluctuated in performance.



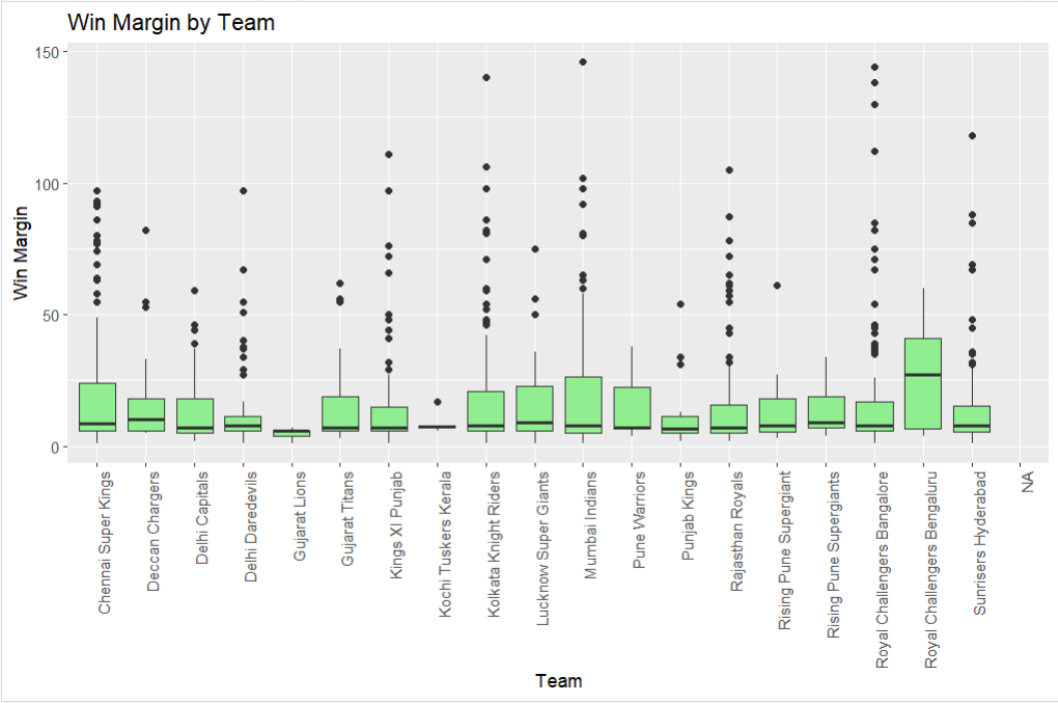
### 4.10 Problem Statement 1: Notable trends in the winning margin by team?

**Approach**: Used boxplots to display the distribution of result_margin for each team.

**Code:**

```
ggplot(ipl_data, aes(x = winner, y = result_margin)) +
geom_boxplot(fill = "lightgreen") +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
labs(title = "Win Margin by Team", x = "Team", y = "Win Margin")
```

**Inference**: Teams like Mumbai Indians and Chennai Super Kings often win by larger margins, reflecting well-rounded performance.




### 4.11 Problem Statement 11: What is the relationship between toss winner and match winner by season?

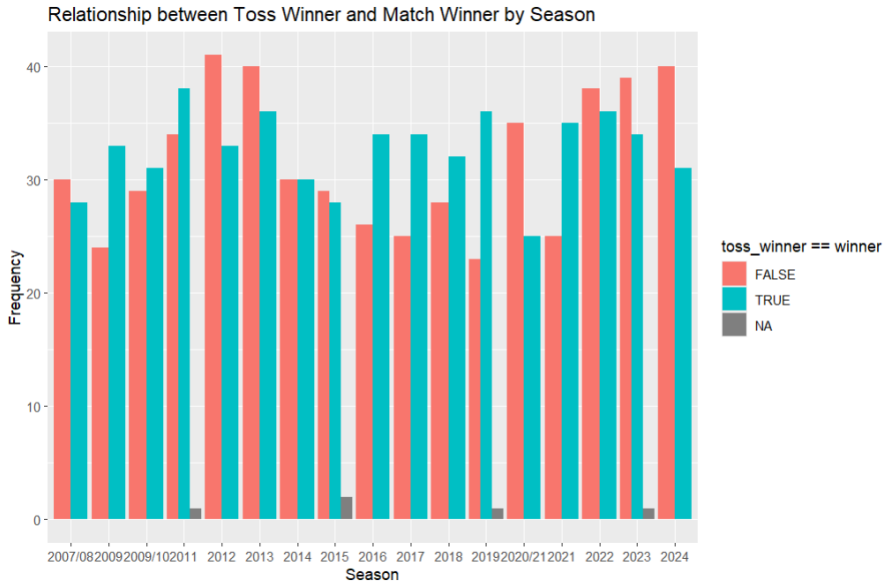
**Approach**: Analyzed whether winning the toss influences match outcomes by season.

**Inference**: Winning the toss can offer an advantage but is not the only factor determining match success.

**Code:**

```
ggplot(ipl_data, aes(x = season, fill = toss_winner == winner)) +
  geom_bar(position = "dodge") +
  labs(title = "Relationship between Toss Winner and Match Winner by Season", x = "Season", y = "Frequency")
```

**Inference**: Winning the toss can offer an advantage but is not the only factor determining match success.




### 4.12 Problem Statement 12: Which team had the highest average win margin over all seasons?

**Approach**: Calculated and visualized each team's average win margin.

**Inference**: Teams with higher average win margins typically demonstrate strong all-round performance.

**Code:**

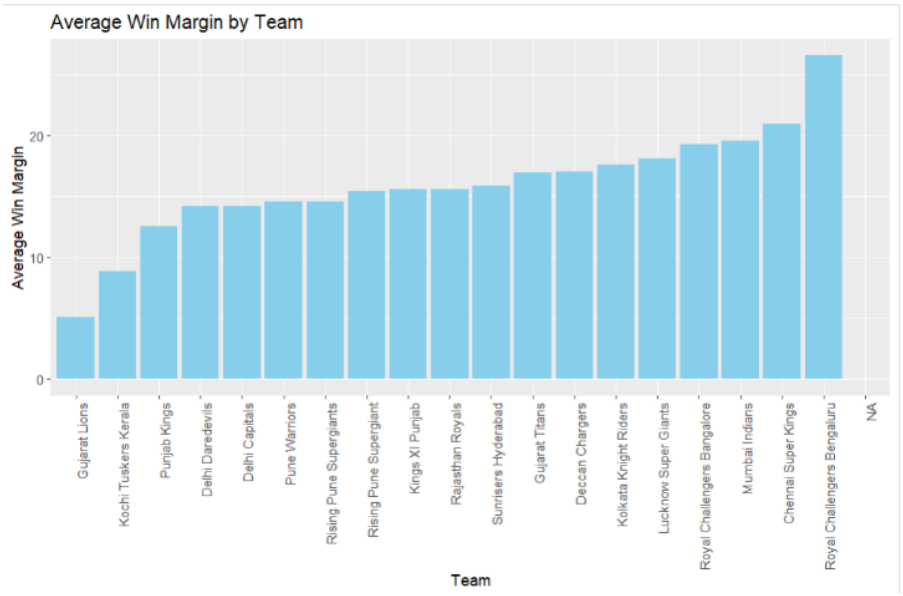
```
avg_win_margin <- ipl_data %>% group_by(winner) %>% summarize(avg_margin = mean(result_margin, na.rm = TRUE))

ggplot(avg_win_margin, aes(x = reorder(winner, avg_margin), y = avg_margin)) +
```

*geom_bar(stat = "identity", fill = "skyblue") +*

*theme(axis.text.x = element_text(angle = 90, hjust = 1)) +*

*labs(title = "Average Win Margin by Team", x = "Team", y = "Average Win Margin")*

**Inference**: Teams with higher average win margins typically demonstrate strong all-round performance.




### 4.13 Problem Statement 13: What is the average win margin by match type?

**Approach**:Compared the average win margin for different match types.

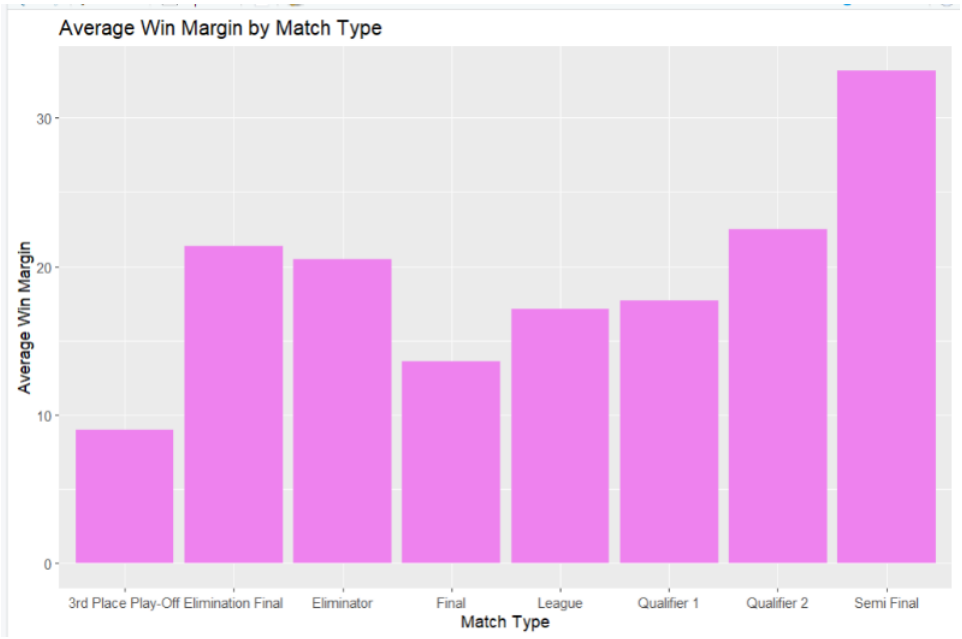
**Code:**

*ggplot(ipl_data, aes(x = match_type, y = result_margin)) +*

*geom_bar(stat = "summary", fun = "mean", fill = "violet") +*

*labs(title = "Average Win Margin by Match Type", x = "Match Type", y = "Average Win Margin")*

**Inference**: Play-offs tend to be closer, while regular season games might have larger victory margins.




## 5. Conclusion

This analysis offers insights into IPL team and player performances, focusing on factors that influence victories and seasonal patterns:

1. **Top Players**: Consistent 'Player of the Match' winners underscore the impact of key players on team success.
2. **Toss Impact**: Winning the toss offers some advantage but does not guarantee a win, highlighting the role of in-game strategies.
3. **Competitive Super Overs**: Teams often involved in super overs show high competitiveness and resilience in close matches.
4. **Venue Influence**: Certain venues yield higher scores due to favorable conditions, while popular venues boost fan engagement.
5. **Seasonal Performance**: Consistent performers, like Mumbai Indians, display adaptability, while run totals fluctuate due to evolving game styles.
6. **Winning Margins**: High target scores often lead to larger win margins, especially for well-rounded teams.
7. **Match Type Variability**: Play-offs are generally closer, reflecting the competitive intensity of high-stakes matches.

These insights highlight how consistency, strategic adaptability, and match context drive IPL success across seasons.