Report: Analysis of the 2021-2022 Soccer Season Data

By Maroof Ansari 20221CSD0030

Data Overview

The dataset soccer21-22.csv contains match data for the 2021-2022 soccer season. The data includes information about match dates, teams, scores, referee involvement, and various match statistics such as shots, fouls, and bookings.

Data Cleaning

1. **Duplicate Removal**: The dataset was cleaned to remove duplicate entries using the distinct() function.
2. **Date Parsing**: Match dates were converted to a proper Date format using the lubridate::dmy() function.
3. **Column Renaming**: Column names were standardized to follow a consistent naming convention using the janitor::clean_names() function

Key Variables Selected

- Match details: date, home_team, away_team, ft_result, and ht_result.
- Performance metrics: Goals, shots, fouls, corners, bookings, and referee data.

Data Summarization

1. **Goals Analysis**:
    - Total goals scored, shots taken, and conversion rates were calculated.
    - A bar chart was created to visualize goals scored by month, showing trends throughout the season.
2. **Referee Performance**:
    - Total referees officiated, fouls committed, and cards issued (yellow and red) were summarized.
    - Visualizations show games officiated and cards issued by each referee.
3. **Team Performance**:
    - Wins, losses, and draws were calculated for each team:

        - **Wins**: Games won by each team.
        - **Losses**: Games lost by each team.
        - **Draws**: Total games ending in a draw.

Visual Insights

1. **Goals by Month**:
    - Monthly trends in goal scoring were highlighted using a column chart. Goals peaked in certain months, possibly due to critical fixtures or player form.
2. **Referee Involvement**:
    - Bar charts showed:
        - Number of games officiated by each referee.
        - Bookings (cards) issued by referees.
        - Red cards issued per referee.
3. **Team Performance**:
    - Bar charts displayed:
        - Total wins and losses per team, emphasizing team performance.
        - Total draws, showing balanced matches between teams.

Statistical Insights

- **Conversion Rate**: Calculated as goals scored divided by total shots, indicating efficiency in converting opportunities into goals.
- **Bookings Analysis**:
    - Total fouls, yellow cards, and red cards provided insights into disciplinary aspects of the matches.

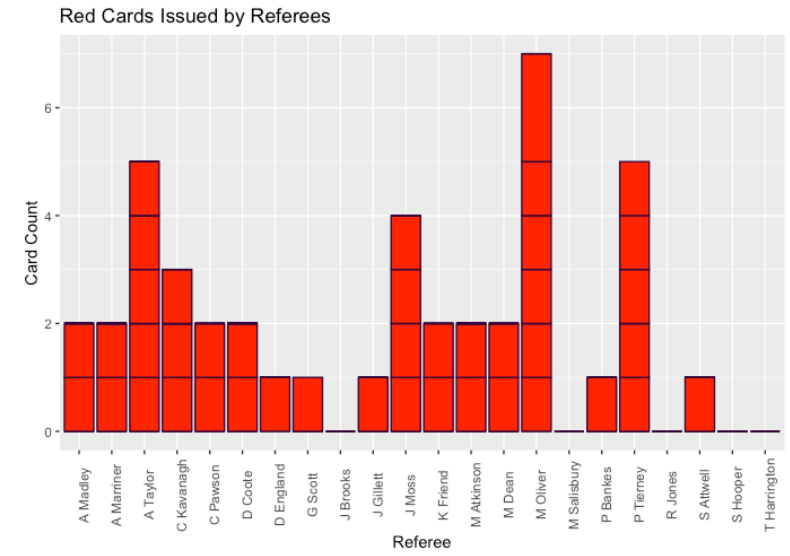
Key Findings

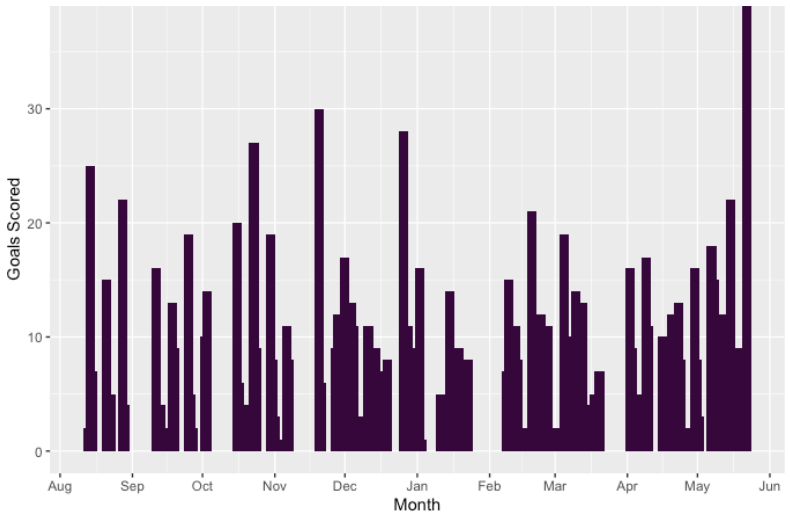
1. **Goal Scoring**: A clear monthly pattern in scoring trends was observed, possibly influenced by team strategies or player conditions.
2. **Referee Impact**: Certain referees were associated with higher booking rates, highlighting strict officiating styles.
3. **Team Dynamics**:
    - Teams with higher wins or losses were identified, reflecting their dominance or struggles in the season.
    - Draws analysis highlighted well-matched teams.

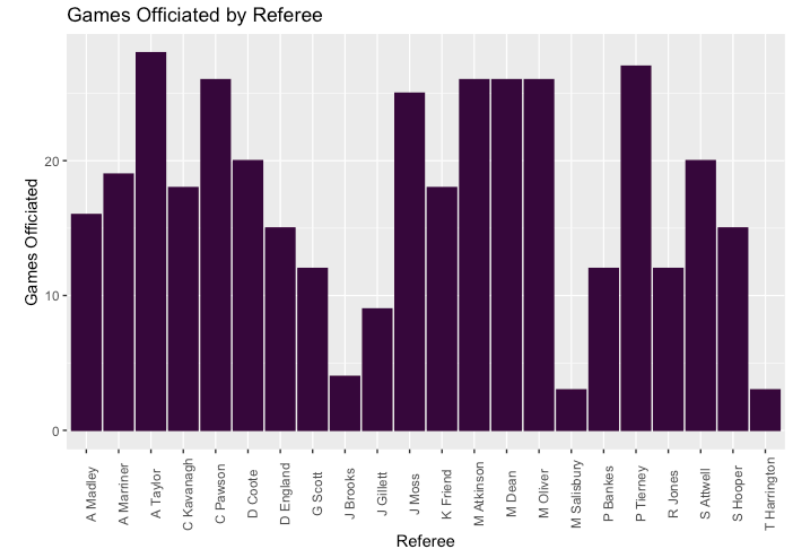
Future Work

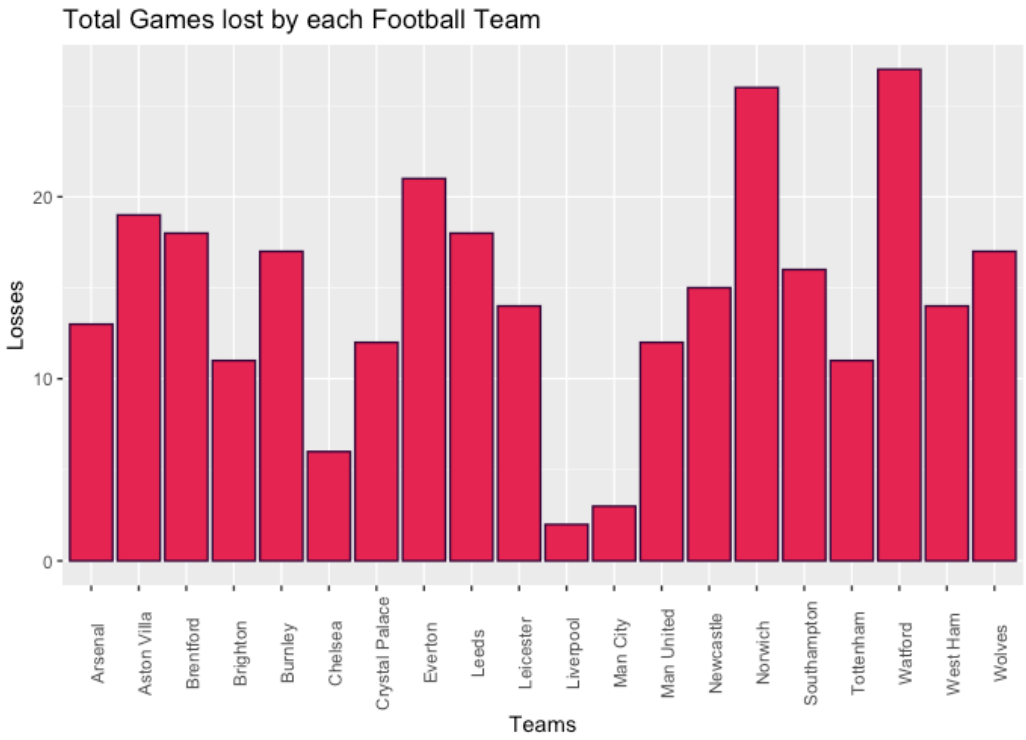
1. Incorporate additional datasets such as player-specific stats or team budgets for a richer analysis.
2. Analyze patterns in home vs. away performance.
3. Extend the visualizations to interactive dashboards for easier exploration.

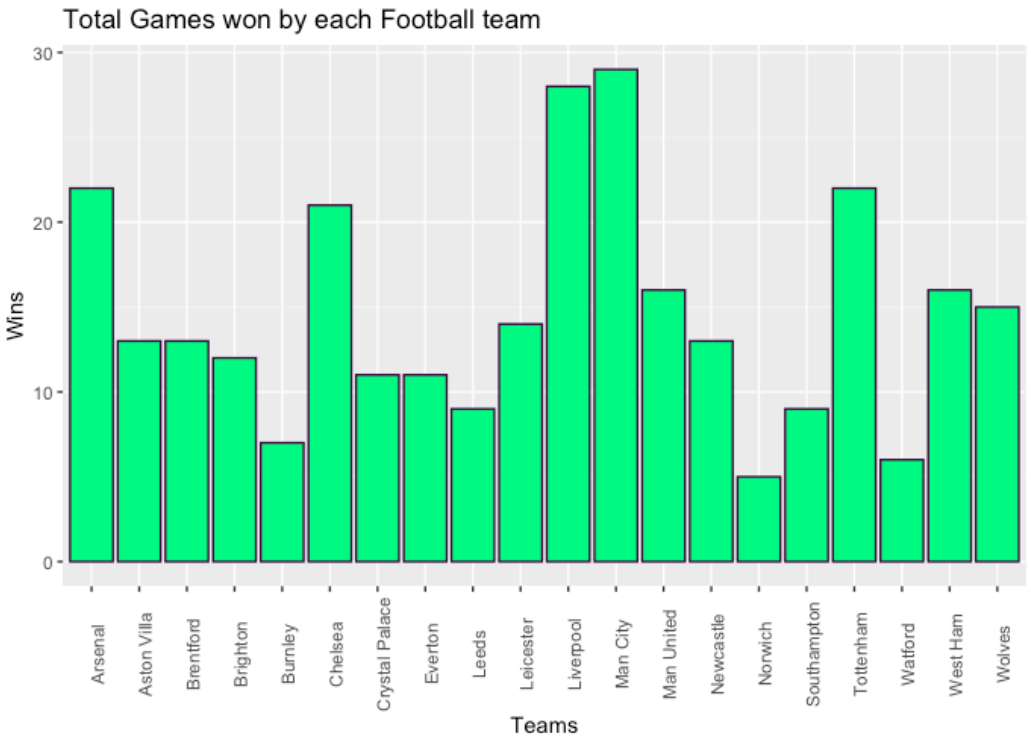
This report highlights the importance of data-driven insights in understanding soccer dynamics, aiding teams, analysts, and enthusiasts in evaluating performance and strategies.

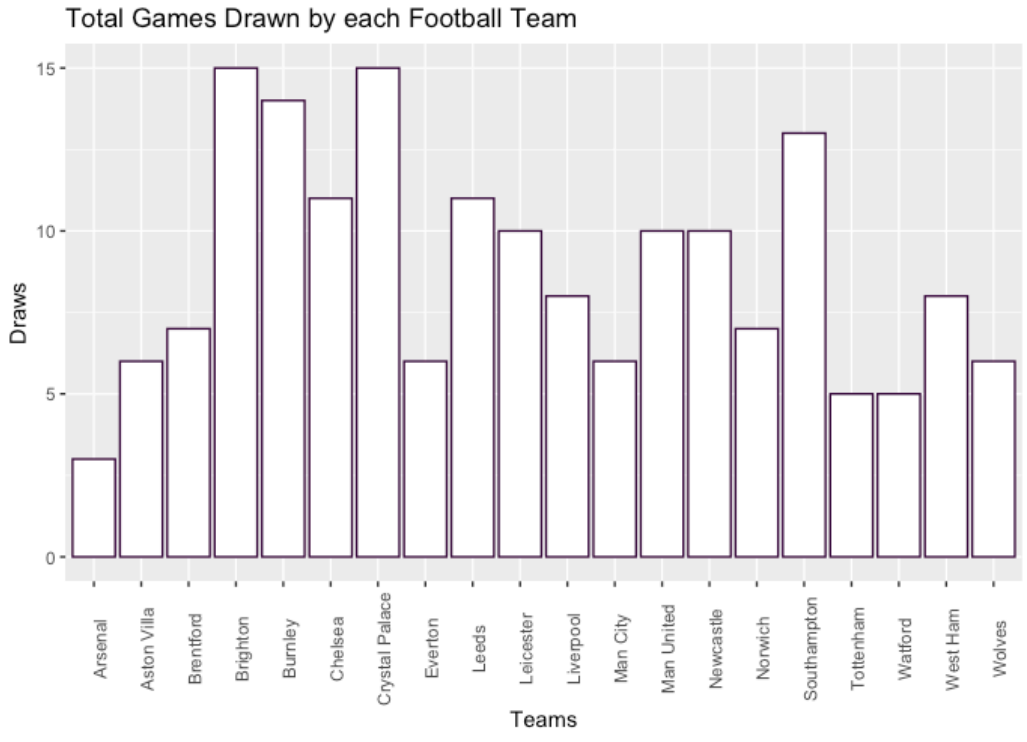







Total Games lost by each Football Team
Losses
0
10
20
Arsenal
Aston Villa
Brentford
Brighton
Burnley
Chelsea
Crystal Palace
Everton
Leeds
Leicester
Liverpool
Man City
Man United
Newcastle
Norwich
Southampton
Tottenham
Watford
West Ham
Wolves
Teams


Total Games won by each Football team
Wins
0
10
20
30
Arsenal
Aston Villa
Brentford
Brighton
Burnley
Chelsea
Crystal Palace
Everton
Leeds
Leicester
Liverpool
Man City
Man United
Newcastle
Norwich
Southampton
Tottenham
Watford
West Ham
Wolves
Teams


Total Games Drawn by each Football Team
Draws
0
5
10
15
Arsenal
Aston Villa
Brentford
Brighton
Burnley
Chelsea
Crystal Palace
Everton
Leeds
Leicester
Liverpool
Man City
Man United
Newcastle
Norwich
Southampton
Tottenham
Watford
West Ham
Wolves
Teams

Code

```
library(tidyverse)
library(janitor)
library(lubridate)

prem_matches <- read.csv('/Users/maroofansari/Downloads/Mini Project /R Project/
soccer21-22.csv')

head(prem_matches)
colnames(prem_matches)

prem_matches <- distinct(prem_matches)
prem_matches$Date <- dmy(prem_matches$Date)
prem_matches <- clean_names(prem_matches)

prem_matches_cleaned <- prem_matches %>%
  select(date, home_team, away_team, ft_home_goals = fthg, ft_away_goals = ftag,
         ft_result = ftr, ht_home_goals = hthg, ht_away_goals = htag, ht_results = htr,
         referee, home_shots = hs, away_shots = as, home_shots_on_target = hst,
         away_shots_on_target = ast, home_fouls = hf, away_fouls = af, home_corners = hc,
         away_corners = ac, home_yellows = hy, away_yellows = ay, home_reds = hr,
away_reds = ar)
head(prem_matches_cleaned)

prem_matches_combined <- prem_matches %>%
  summarise(date, home_team, away_team, goals_scored = fthg + ftag, result = ftr,
            referee, total_shots = hs + as, shots_on_target = hst + ast,
            fouls_committed = hf + af, corner_count = hc + ac, bookings = hy+ay+hr+ar,
yellow_cards = hy+ay, red_cards = hr+ar)
head(prem_matches_combined)

distinct(prem_matches,football_clubs = home_team) %>%
  arrange(football_clubs)

prem_matches_combined %>%
  summarise(total_goals = sum(goals_scored),total_shots = sum(total_shots),
            total_shots_on_target = sum(shots_on_target),
            conversion_rate = total_goals/total_shots*100)

ggplot(data = prem_matches_combined, aes(x = date, y = goals_scored)) +
```

```
  geom_col(width = 4, fill = "#38003c") +
  scale_x_date(date_labels="%b",date_breaks  ="1 month") +
  xlab("Month") + ylab("Goals Scored") + labs(title = "Goals by Month")

prem_matches_combined %>%
  summarise(referees_used = sum(count(distinct(prem_matches, referee))),
            total_fouls = sum(fouls_committed),
            total_bookings = sum(bookings),
            total_yellow_cards = sum(yellow_cards),
            total_red_cards = sum(red_cards))

ggplot(data = prem_matches_combined, aes(x = referee)) +
  geom_bar(fill = "#38003c", color = "#38003c") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Referee") + ylab("Games Officiated") +
  labs(title = "Games Officiated by Referee")

ggplot(data = prem_matches_combined, aes(x = referee, y = bookings, fill = bookings)) +
  geom_col(fill = "#38003c", color = "#38003c") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Referee") + ylab("Cards Issued") +
  labs(title = "Bookings Issued by Referees")

ggplot(data = prem_matches_combined, aes(x = referee, y = red_cards, fill = bookings)) +
  geom_col(fill = "red", color = "#38003c") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Referee") + ylab("Card Count") +
  labs(title = "Red Cards Issued by Referees")

wins <- prem_matches_cleaned %>%
  mutate(ft_result = case_when(
    ft_result == "H" ~ home_team,
    ft_result == "A" ~ away_team,
    ft_result == "D" ~ "Draw"))
wins <- wins %>% filter(ft_result != "Draw")

ggplot(data = wins, aes(x = ft_result)) +
  geom_bar(fill = "#00ff85", color = "#38003c") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Teams") + ylab("Wins") + labs(title = "Total Games won by each Football team")
```

```
losses <- prem_matches_cleaned %>%
  mutate(ft_result = case_when(
    ft_result == "H" ~ away_team,
    ft_result == "A" ~ home_team,
    ft_result == "D" ~ "Draw"))
losses <- losses %>% filter(ft_result != "Draw")

ggplot(data = losses, aes(x = ft_result)) +
  geom_bar(fill = "#e90052", color = "#38003c") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Teams") + ylab("Losses") + labs(title = "Total Games lost by each Football Team")

draws <- prem_matches_cleaned %>%
  select(home_team,away_team,ft_result) %>%
  mutate(ft_result = case_when(
    ft_result == "H" ~ home_team,
    ft_result == "A" ~ away_team,
    ft_result == "D" ~ "Draw"))
draws <- draws %>% filter(ft_result == "Draw")
draws <- cbind(draws[3], stack(draws[1:2]))

ggplot(data = draws, aes(x = values)) +
  geom_bar(fill = "#ffffff", color = "#38003c") +
  theme(axis.text.x = element_text(angle = 90)) +
  xlab("Teams") + ylab("Draws") + labs(title = "Total Games Drawn by each Football
Team")
```