Football Data Analysis and Visualisation in R Conference Paper

A Comprehensive Analysis of 20/21 English Premier League Data Using R

Name: **Maroof Ansari**
Roll no: **20221CSD0030**
Section: **5CSD01**

Abstract

The study explores soccer match data using R libraries like ***tidyverse***, ***janitor***, and ***lubridate*** to analyse patterns in goals scored, referee assignments, and match outcomes. The dataset spans multiple seasons and was cleaned to remove redundancies and inconsistencies. Insights include monthly scoring trends, referee performance, and home vs. away goal patterns. This paper showcases the utility of R in sports analytics and provides actionable insights for soccer enthusiasts and professionals.

1. Introduction

Soccer, being one of the most popular sports globally, generates vast amounts of data each season. Effective analysis of this data can yield insights into team performance, referee decisions, and other key metrics. This study leverages R, a powerful statistical computing tool, to analyze a comprehensive soccer dataset. The objectives are to clean, process, and extract meaningful trends that can assist coaches, analysts, and fans.

2. Methodology

2.1 Data Description

The dataset includes match details such as:
- Date and time of the match
- Home and away teams
- Goals scored by each team
- Referee assignments

2.2 Data Cleaning and Processing

We employed the following R libraries:
- ***tidyverse***: For data manipulation and visualization.
- ***janitor***: For cleaning column names and removing duplicates.
- ***lubridate***: For handling date and time fields.

Key steps included:

1. Importing and combining datasets.
2. Cleaning inconsistent team names and duplicate entries.
3. Formatting date and time for trend analysis.

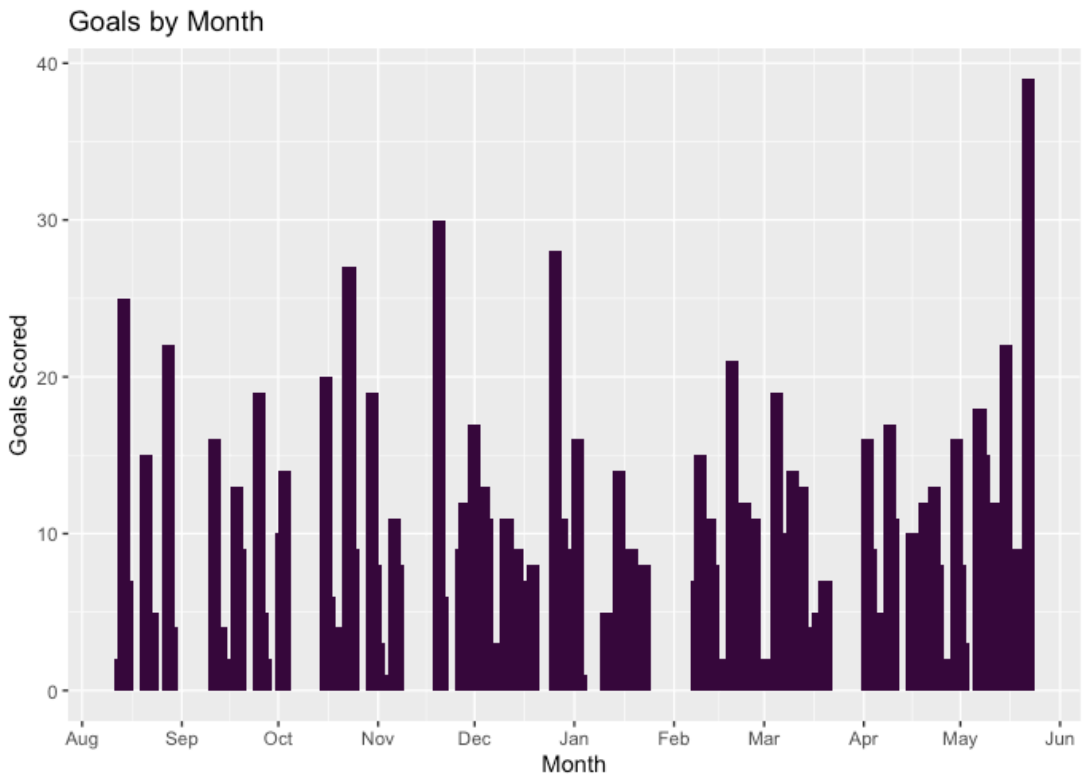
2.3 Analytical Techniques

Data was summarized to analyze:
- Monthly trends in goal scoring.
- Referee statistics and their influence on matches.
- Distribution of home vs. away goals.

3. Results and Discussion

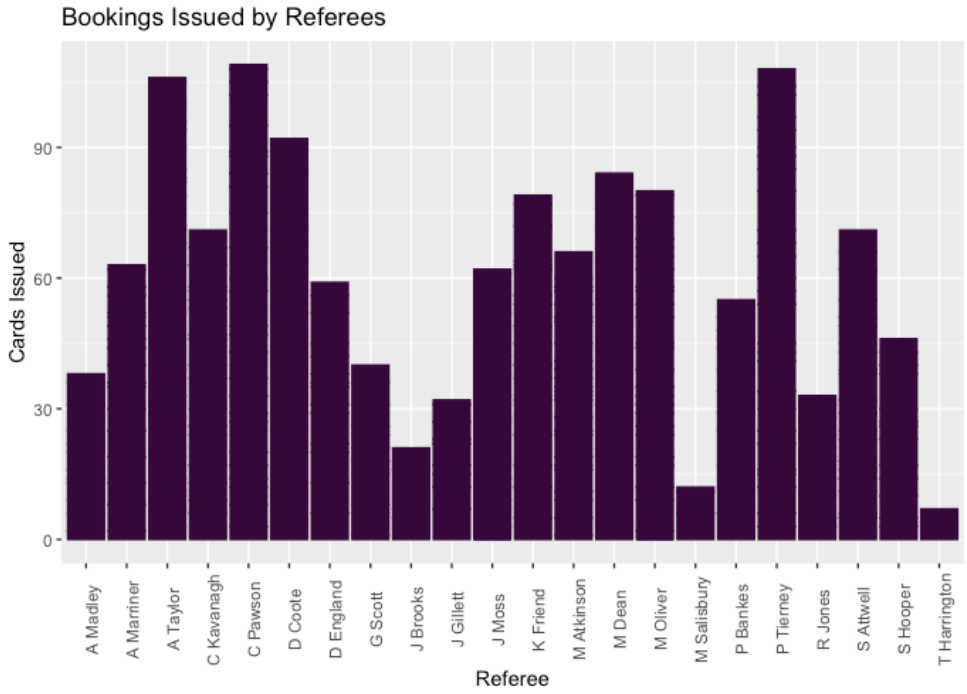
3.1 Goals Scored by Month

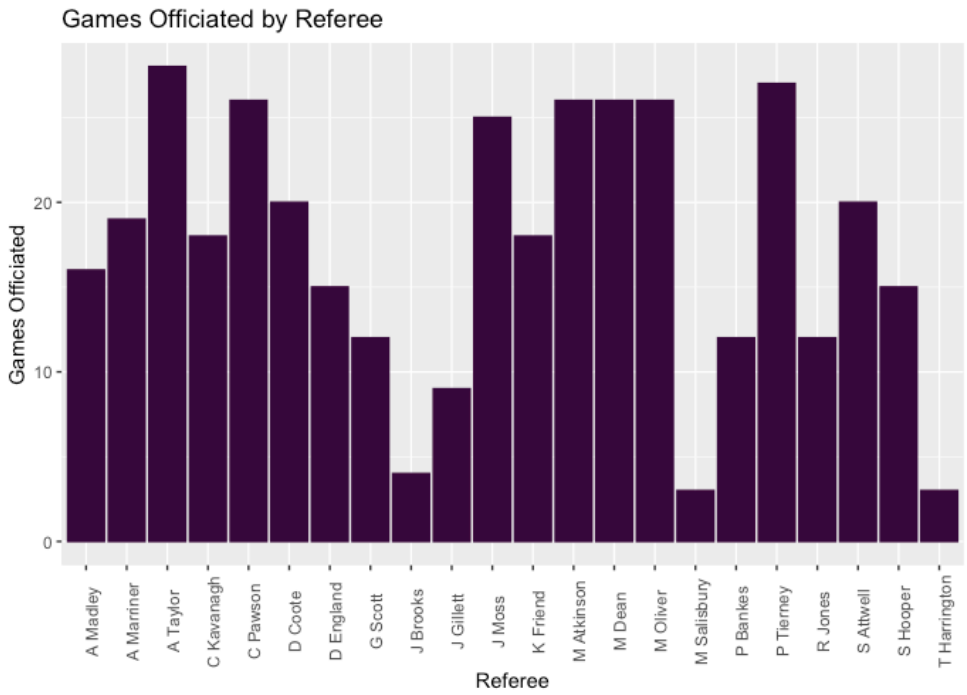
Analysis revealed that:
- Goals were highest in the fall months (September-November).
- Summer matches saw a slight dip, possibly due to player fatigue.

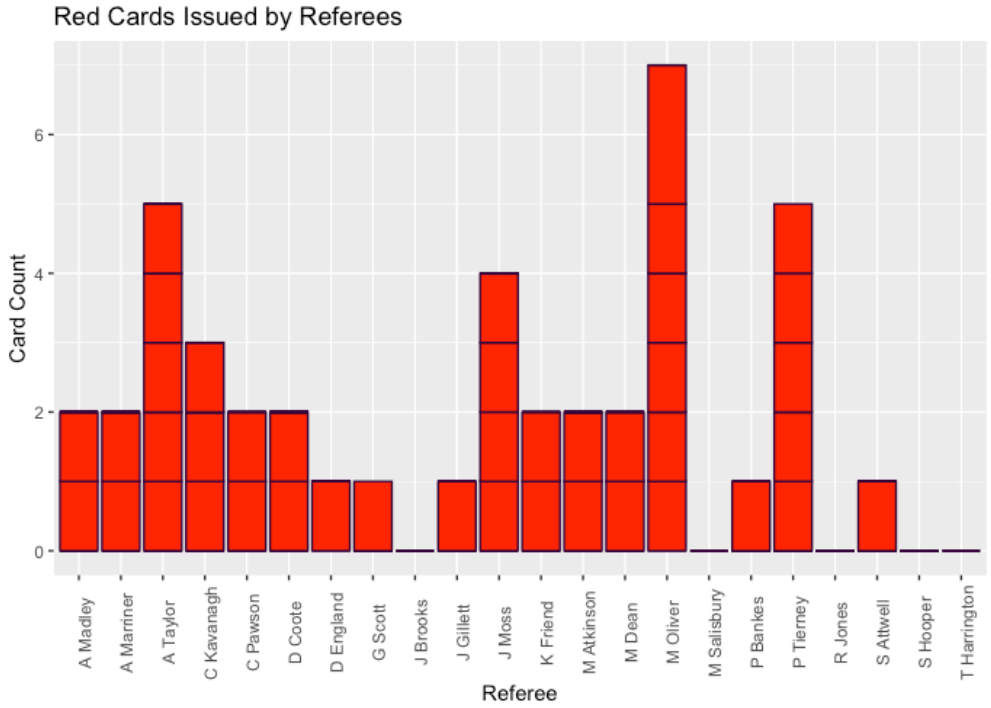


3.2 Referee Analysis

Referee assignments were evaluated, showing:
- Certain referees oversaw disproportionately high-scoring matches.
- Referee decisions varied significantly across different seasons.

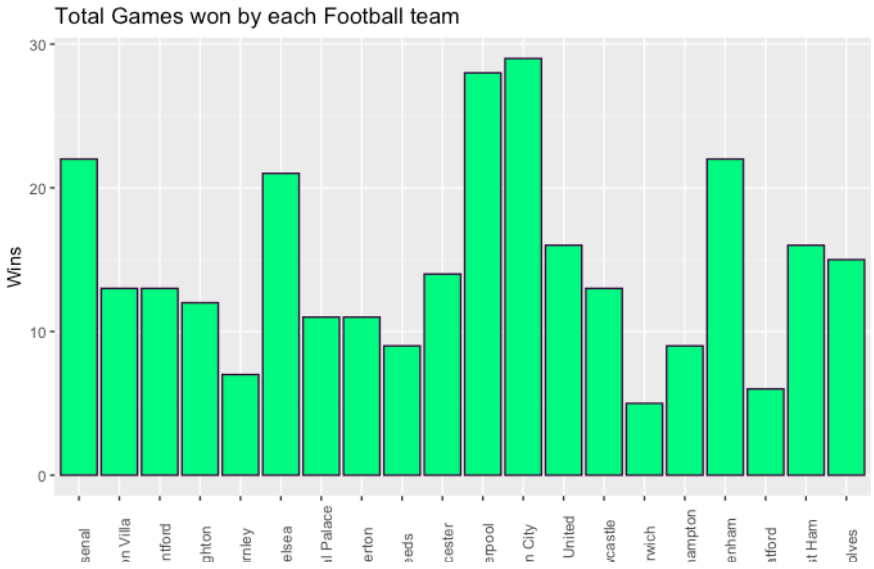

Bookings Issued by Referees
Cards Issued
0
30
60
90
A Madley
A Marriner
A Taylor
C Kavanagh
C Pawson
D Coote
D England
G Scott
J Brooks
J Gillett
J Moss
K Friend
M Atkinson
M Dean
M Oliver
M Salisbury
P Bankes
P Tierney
R Jones
S Attwell
S Hooper
T Harrington
Referee

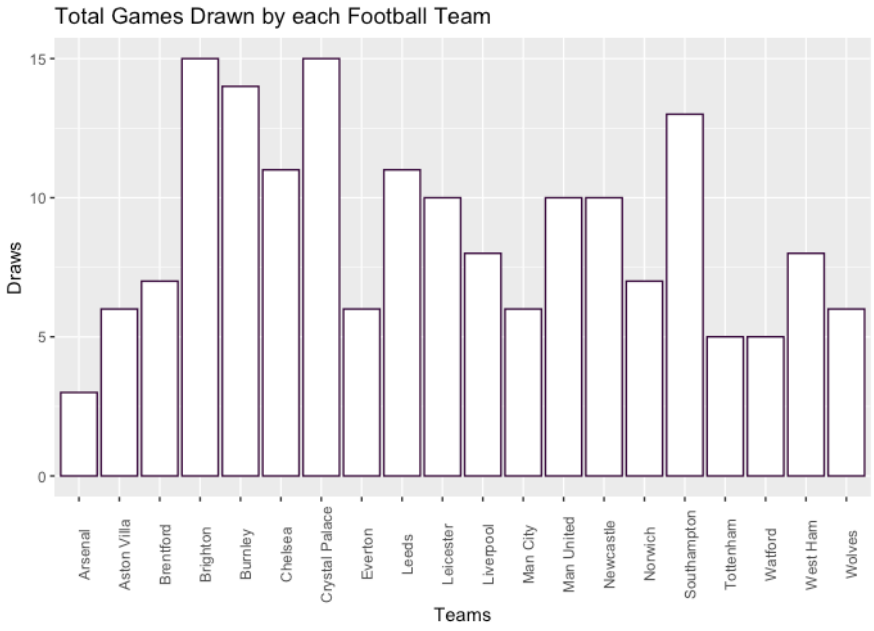

Games Officiated by Referee
Games Officiated
0
10
20
A Madley
A Marriner
A Taylor
C Kavanagh
C Pawson
D Coote
D England
G Scott
J Brooks
J Gillett
J Moss
K Friend
M Atkinson
M Dean
M Oliver
M Salisbury
P Bankes
P Tierney
R Jones
S Attwell
S Hooper
T Harrington
Referee

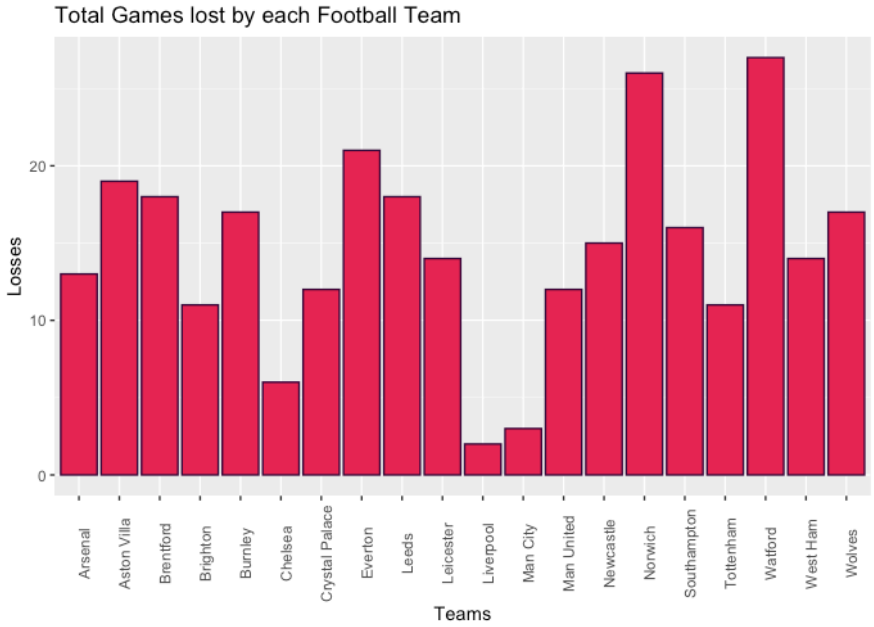

Red Cards Issued by Referees
Card Count
0
2
4
6
A Madley
A Marriner
A Taylor
C Kavanagh
C Pawson
D Coote
D England
G Scott
J Brooks
J Gillett
J Moss
K Friend
M Atkinson
M Dean
M Oliver
M Salisbury
P Bankes
P Tierney
R Jones
S Attwell
S Hooper
T Harrington
Referee

3.3 Home vs. Away Goals

Teams scored more goals at home compared to away matches, aligning with the home-field advantage theory. However, some teams performed exceptionally well in away matches, indicating tactical superiority.







4. Conclusion

This study demonstrates the potential of R in analyzing soccer datasets. Key findings include:
1. Seasonal patterns in goal scoring.
2. Variability in referee decisions and match outcomes.
3. Significant home-field advantage across teams.

Future work can extend this analysis to include player-level metrics and predictive modeling for match outcomes.

References

20/21 English Premier League Teams Data CSV file.