Project Report

Vuk Radulovic

0.1 Introduction

0.1.1 Research questions:

1) How has the **average number of goals per international match** changed over time?
2) How strong is **home field advantage**, and does it differ between **friendly games** and **tournaments**?

I will use historical match dataset (1872-present), and from it, use year, total goals, result towards answering the research questions. Additionally, I will use a compiled **confederation** mapping, and a basic **Elo-style rating** to add context to the overall result.

0.2 Background Information

Soccer has been around for centuries, and with it, the state of the game changed. More notably, the strategies developed, players becoming more athletic, etc. These naturally affect the scores in games, which is what we're looking to research. Additionally, within soccer community, home field advantage is known to affect the state of the game, mostly due to the atmosphere home crowd is able to create, and this is what will look for in our data as well. Lastly, for further context we will look into friendly games vs tournaments, and official team rankings, as these will add context into scores and home field advantage.

0.3 Data Summary

Primary: International football results, 1872-present (each row is one match, fields include date, teams, scores, tournament, location) https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017

Secondary A (compiled): `country_confederation.csv`- Compiled mapping(team -> UEFA/CONMEBOL/CONCACAF/AFC/CAF/OFC)

Table 1: Matches, average goals, and draw rate by decade

```
# A tibble: 16 x 4
   decade n_matches avg_goals draw_rate
   <chr>      <int>     <dbl>     <dbl>
 1 1870s         13      4.54    0.154
 2 1880s         52      5.71    0.0962
 3 1890s         60      5.18    0.15
 4 1900s        150      4.79    0.167
 5 1910s        345      4.08    0.177
 6 1920s        869      3.89    0.188
 7 1930s       1168      4.20    0.167
 8 1940s        890      4.30    0.157
 9 1950s       1874      4.03    0.178
10 1960s       3345      3.52    0.193
11 1970s       4686      2.98    0.214
12 1980s       5650      2.54    0.259
13 1990s       7055      2.75    0.248
14 2000s       9508      2.77    0.236
15 2010s       9709      2.66    0.238
16 2020s       5284      2.67    0.235
```

Secondary B (derived): `team_strength_ratings.csv`- Computed simple Elo-style rating from the primary data (everyone starts at 1500; K=20,draw=0.5, no home bonus)

countries_names.csv: Mapping to standardize name changes over the years.

0.3.1 Attributes used

I will focus on: `year`, `total_goals`,`result`,`match_type` (friendly vs tournament), `neutral`, `home_confed`/`away_confed`, and `home_rating`/`away_rating`

0.4 EDA

0.4.1 Tables

Table 1 shows match count, average goals, and draw rates by decade.

We can see a trend of reduction in amount of average goals

?@tbl-topten shows top 10 highest scoring international matches

Table 2: Results by match type(friendly vs tournament)

match_type	home_win_rate	draw_rate	away_win_rate	n
Friendly	0.516	0.253	0.231	19217
Tournament	0.611	0.214	0.175	31441

date	home_team	away_team	home_score	away_score	total_goals
2001-04-11	Australia	Eastern Samoa	31	0	31
1971-09-13	Tahiti	Cook Islands	30	0	30
1979-08-30	Fiji	Kiribati	24	0	24
2001-04-09	Australia	Tonga	22	0	22
1966-04-01	Libya	Oman	21	0	21
2005-03-11	North Korea	Guam	21	0	21
1987-12-15	Papua New Guinea	Eastern Samoa	20	0	20
2000-02-14	Kuwait	Bhutan	20	0	20
1983-08-22	Papua New Guinea	Niue	19	0	19
2000-01-26	China	Guam	19	0	19

These will most likely be outliers given the sheer amount of goals scored

Table 2 compares result rates by match type- friendlies vs tournament

In both cases we see a higher rate of home wins, with tournament games having higher home win rate and lower away win rate

0.4.2 Visualization

?@fig-goaldist shows distribution of total goals

As we can see, most games have about 2-3 goals

Figure 2 shows the overall outcomes in Home/Draw/Away games

Graph shows us home teams win significantly more games

Figure 3 shows average goals over time. This will show trends and changes in it

The trend shows a significant drop in goals scored across time

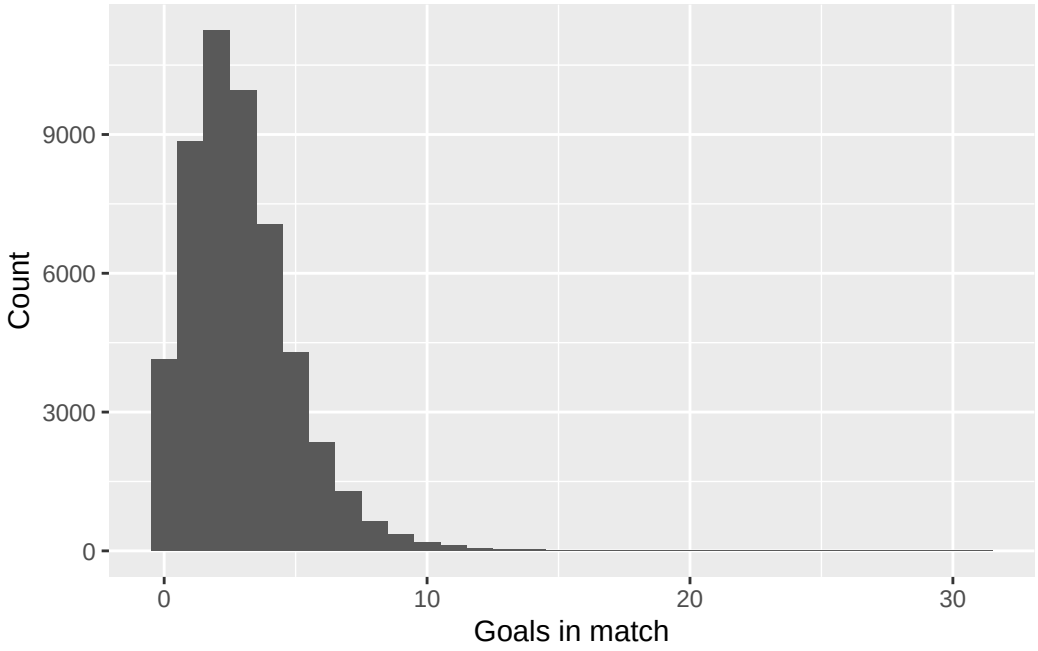

Figure 1: Distribution of total goals per match

Figure 4 shows home advantage with rating gap added into context, this allows us to better understand home wins

The graph tells us the higher the rating gap is for the home team, the more they win their home games

0.5 Conclusion

Looking at all the available data, we can draw a good number of conclusions. Firstly, there is a significant reduction in average goals since th 1870s. Back then, they averaged 4.54 goals while in the 2020s they averaged 2.67 goals. It is important to note:

1. The draw rate is higher in 2020s vs 1870s
2. There have been only 13 matches in 1870s vs 2020s

This is significant because the data is skewed because the volume of games increased across the years. Next, when looking at friendly vs tournament games, we see a higher rate of home wins during tournaments compared to friendlies. It is important to note that there are almost double the amount of tournament games compared to friendlies which might skew results. Furthermore, we can see that teams with a higher rating gap will win home games at a higher

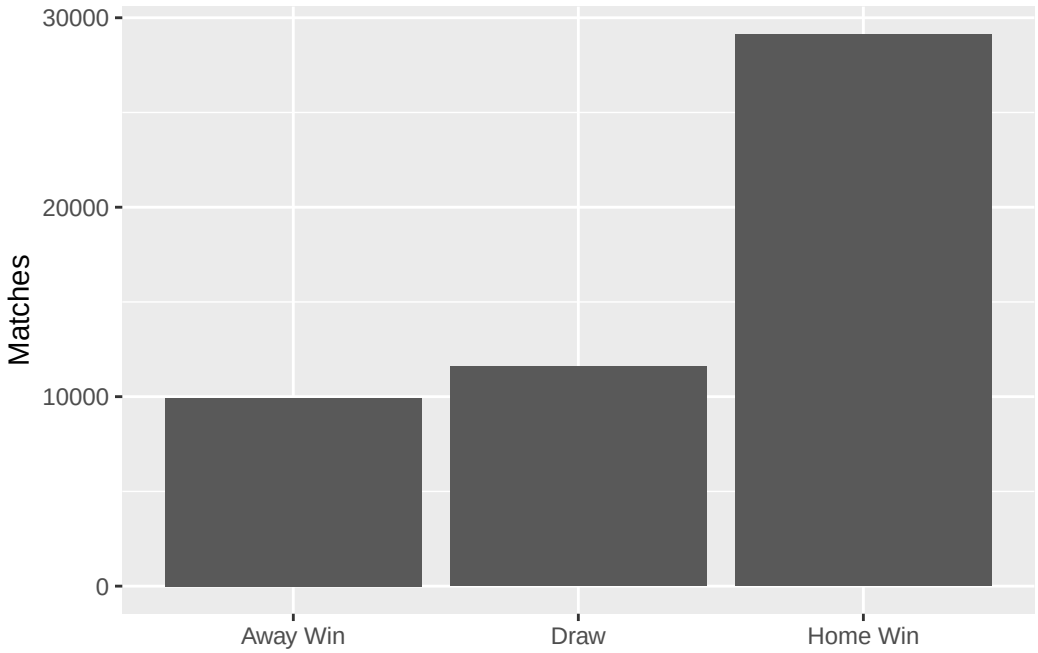

Figure 2: Overall result breakdown

rate, as seen in Figure 4. This could be supported by the table showing top 10 highest scoring matches.

The biggest limitation is lack of more data, we are working with data that will give us a glimpse into our research question. With more data points we can introduce more context into the research. For example, we know there were only 13 games played in 1870s, and we're halfway into 2020s so we can't compare 2020s with 2010s that have almost double the amount of games. Further steps for this research would be more in depth look into each decade.

0.6 References

International football results(Jürisoo, Mart. "International Football Results from 1872 to 2025." Kaggle, 7 July 2025, www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017. Accessed 13 Aug. 2025.)

Author-compiled `country_confederation.csv`("Inside FIFA." FIFA, inside.fifa.com/associations. Accessed 13 Aug. 2025.)

Author-compiled `team_strength_ratings.csv` (method: simple Elo update from primary data)

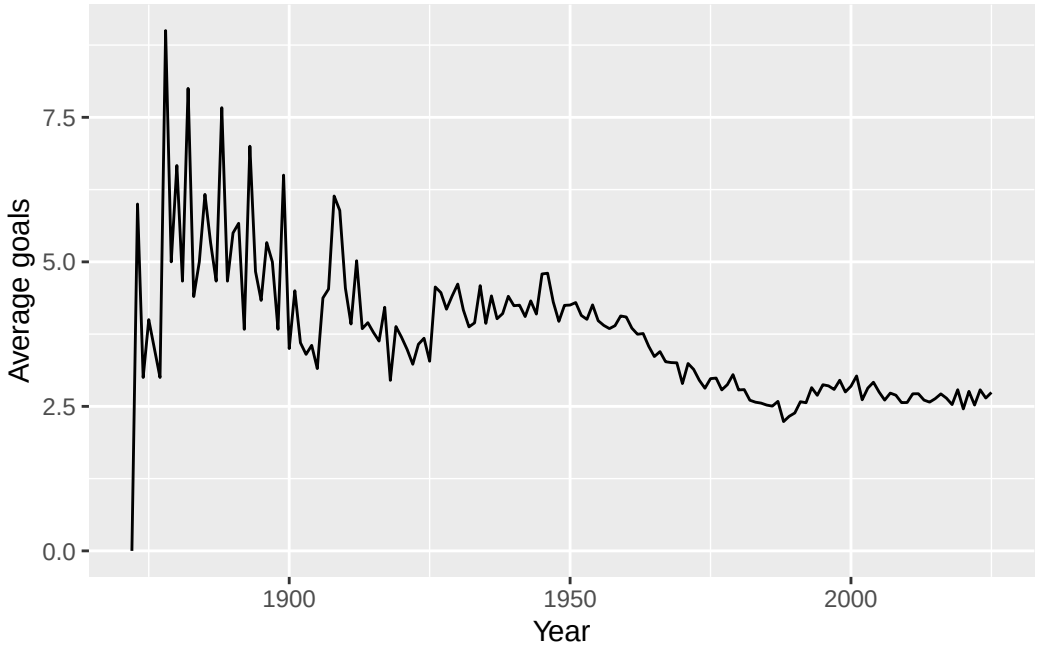

Figure 3: Average goals per match by year

R for Data science (Hadley Wickham, Mine Çetinkaya-Rundel. "R for Data Science (2E)." R for Data Science (2e), r4ds.hadley.nz/. Accessed 13 Aug. 2025.)

0.7 Code Appendix

```
library(tidyverse)
library(lubridate)
library(stringr)

data_dir <- "Stat184DATA"
matches <- readr::read_csv("C:/Users/Jake Iovacchini/OneDrive/Desktop/Stat184DATA/all_matches
confed <- readr::read_csv("C:/Users/Jake Iovacchini/OneDrive/Desktop/Stat184DATA/country_con
ratings <- readr::read_csv("C:/Users/Jake Iovacchini/OneDrive/Desktop/Stat184DATA/team_streng
countries <- readr::read_csv("C:/Users/Jake Iovacchini/OneDrive/Desktop/Stat184DATA/countries

# Wrangling
matches_clean <- matches |>
  mutate(
    date = ymd(date),
```

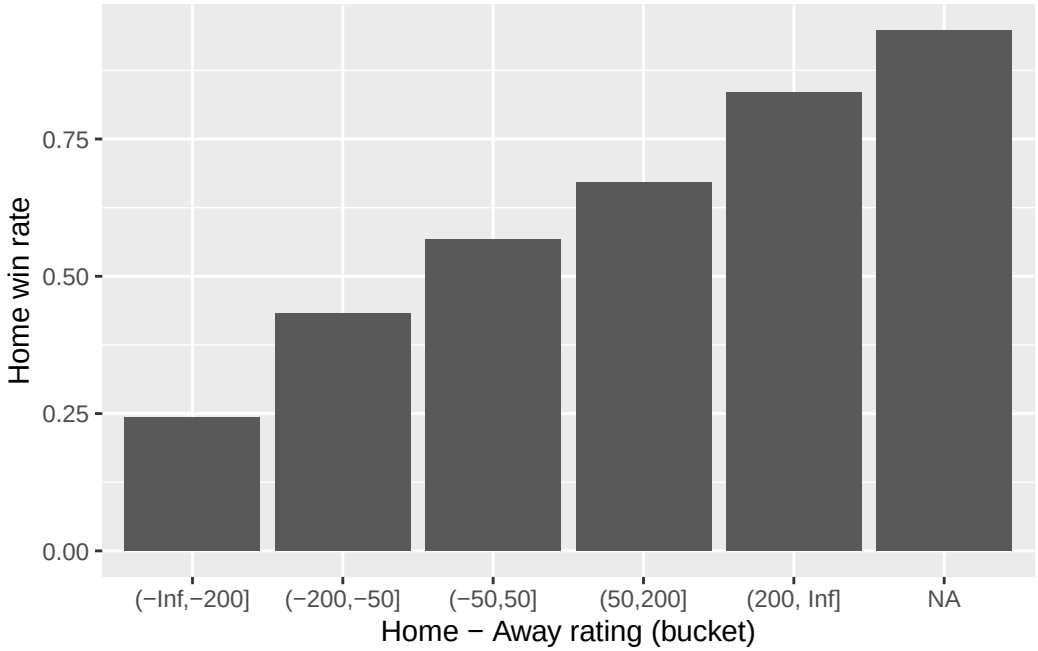

Figure 4: Home win rate vs rating gap (bucketed)

```
    year = year(date),
    total_goals = home_score + away_score,
    result = case_when(
      home_score > away_score ~ "Home Win",
      home_score < away_score ~ "Away Win",
      TRUE ~ "Draw"
    )
  )
# Normalization
clean_tournament <- function(x) {
  x |>
    str_to_lower() |>  # making everything lowercase
    str_replace_all("\\s+", " ") |> # making everything single space
    str_replace_all("qualif(ication)?s?", "qualifiers") |> # "qualifications" turned into "qu
    str_replace_all("fifa world cup( finals)?", "world cup") |> #normalizing into world cup
    str_trim()
}

matches_std <- matches_clean |>
  mutate(
```

```
    tourn_clean = clean_tournament(tournament), # make naming consistent
    match_type  = if_else(str_detect(tourn_clean, "friendly"), "Friendly", "Tournament"), # c
    home_team_std = if (!is.null(countries))
      coalesce(recode(home_team, !!!setNames(countries$current_name, countries$original_name)
    away_team_std = if (!is.null(countries))
      coalesce(recode(away_team, !!!setNames(countries$current_name, countries$original_name)
  )

# We'll be joining confederations with ratings
matches_confed <- matches_std |>
  left_join(confed, by = c("home_team_std" = "team_name")) |>
  rename(home_confed = confederation) |>
  left_join(confed, by = c("away_team_std" = "team_name")) |>
  rename(away_confed = confederation)

matches_feat <- matches_confed |>
  left_join(ratings |> select(team, rating) |> rename(home_team_std = team, home_rating = rat
            by = "home_team_std") |>
  left_join(ratings |> select(team, rating) |> rename(away_team_std = team, away_rating = rat
            by = "away_team_std")

# pivot_wider option
yearly_counts_wide <- matches_feat |>
  count(year, result) |>
  tidyr::pivot_wider(names_from = result, values_from = n, values_fill = 0) |>
  arrange(year)

# Reduction
by_decade <- matches_feat |>
  mutate(decade = paste0(floor(year/10)*10, "s")) |>
  group_by(decade) |>
  summarise(n_matches = n(),
            avg_goals = mean(total_goals, na.rm = TRUE),
            draw_rate = mean(result == "Draw"),
            .groups = "drop")

top_scoring <- matches_feat |>
  arrange(desc(total_goals)) |>
  select(date, home_team, away_team, home_score, away_score, total_goals) |>
  slice_head(n = 10)

by_type <- matches_feat |>
```

```
  group_by(match_type) |>
  summarise(home_win_rate = mean(result == "Home Win"),
            draw_rate = mean(result == "Draw"),
            away_win_rate = mean(result == "Away Win"),
            n = dplyr::n(), .groups = "drop")

by_neutral <- matches_feat |>
  group_by(neutral) |>
  summarise(home_win_rate = mean(result == "Home Win"),
            draw_rate = mean(result == "Draw"),
            away_win_rate = mean(result == "Away Win"),
            n = dplyr::n(), .groups = "drop")

rating_gap_tbl <- matches_feat |>
  mutate(rating_gap = home_rating - away_rating,
         gap_bucket = cut(rating_gap, breaks = c(-Inf,-200,-50,50,200,Inf))) |>
  group_by(gap_bucket) |>
  summarise(home_win_rate = mean(result == "Home Win"),
            n = dplyr::n(), .groups = "drop")
by_decade
top_scoring
by_type
ggplot(matches_feat, aes(total_goals)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Goals in match", y = "Count")
ggplot(matches_feat, aes(result)) +
  geom_bar() +
  labs(x = NULL, y = "Matches")
matches_feat |>
  group_by(year) |>
  summarise(avg_goals = mean(total_goals, na.rm = TRUE), .groups = "drop") |>
  ggplot(aes(year, avg_goals)) +
  geom_line() +
  labs(x = "Year", y = "Average goals")
rating_gap_tbl |>
  ggplot(aes(gap_bucket, home_win_rate)) +
  geom_col() +
  labs(x = "Home - Away rating (bucket)", y = "Home win rate")
```