New EDA

Team 23

2025-12-05

Cleaning qualified_teams.csv and ranking_soccer_1901-2023.csv datasets

```
elo_wc <- add_wc_matches_to_elo(
  elo_filepath = "data/national_ranking_soccer_1901-2023.csv",
  wc_perf_filepath = "data/qualified_teams.csv"
)

head(elo_wc)
```

```
## # A tibble: 6 x 8
##   world_cup_year team_name   elo_rating elo_rank elo_1yr_change_rating win_ratio
##            <dbl> <chr>            <dbl>    <dbl>                  <dbl>     <dbl>
## 1           1982 West Germa~       2111        1                    -12     0.566
## 2           1982 Brazil            2091        2                     46     0.627
## 3           1982 Argentina         2006        3                      7     0.538
## 4           1982 Soviet Uni~       2001        4                     49     0.557
## 5           1982 Poland            1959        5                     12     0.425
## 6           1982 Belgium           1917        6                    -36     0.386
## # i 2 more variables: goals_ratio <dbl>, count_matches <dbl>
```

```
write_csv(elo_wc, "data/primary_dataset_wc_elo.csv")
```

```
source("02_funct_wc_elo_eda.R")
wc <- load_wc_elo()
head(wc)
```

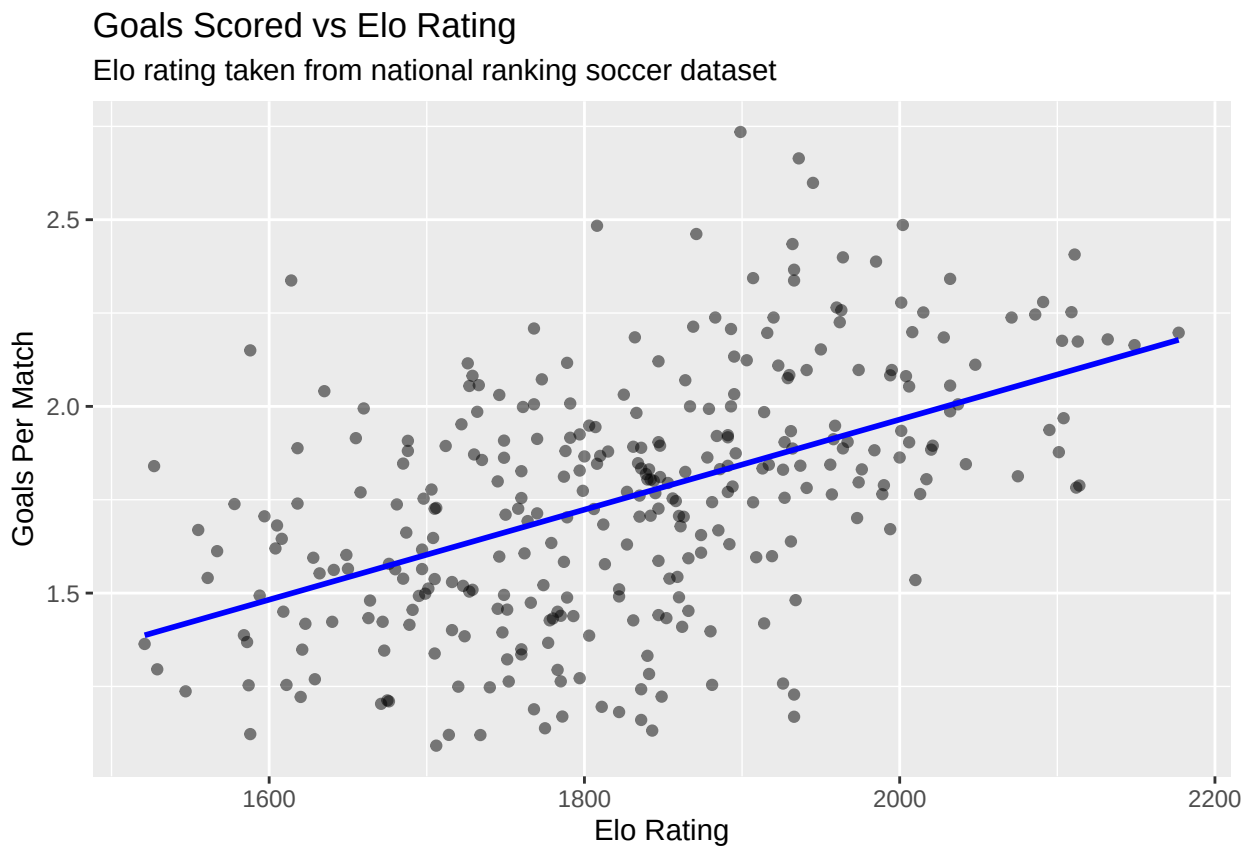
```
## # A tibble: 6 x 8
##   world_cup_year team_name   elo_rating elo_rank elo_1yr_change_rating win_ratio
##            <dbl> <chr>            <dbl>    <dbl>                  <dbl>     <dbl>
## 1           1982 West Germa~       2111        1                    -12     0.566
## 2           1982 Brazil            2091        2                     46     0.627
## 3           1982 Argentina         2006        3                      7     0.538
## 4           1982 Soviet Uni~       2001        4                     49     0.557
## 5           1982 Poland            1959        5                     12     0.425
## 6           1982 Belgium           1917        6                    -36     0.386
## # i 2 more variables: goals_ratio <dbl>, count_matches <dbl>
```

1) Goals vs. Elo Ratings

We first explored whether team strength (as measured by Elo rating from the national ranking dataset) helps explain how many goals a team scores in a World Cup tournament.

```
plot_goals_vs_elo(wc)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



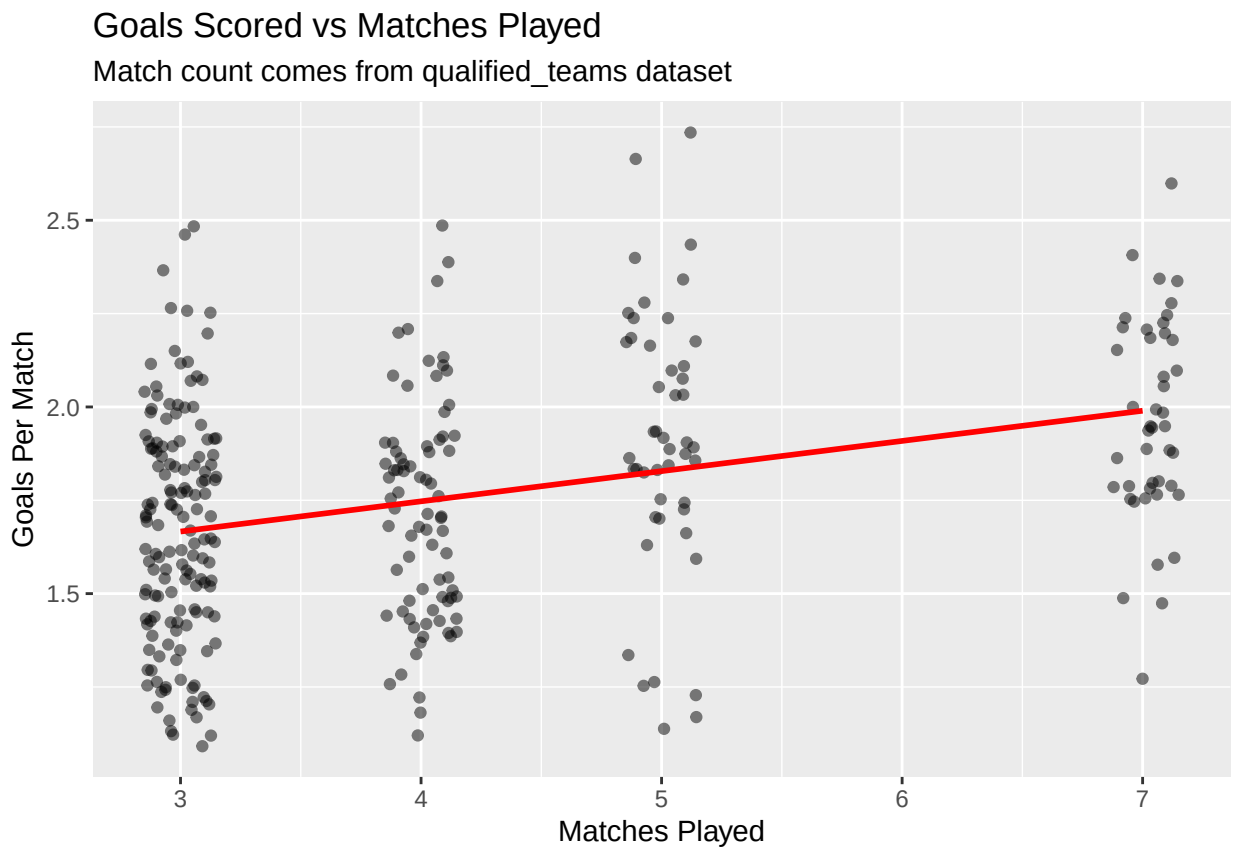
This scatterplot shows a clear, positive relationship between team Elo ratings and the number of goals they score per match in the World Cup. While there is noticeable variability, teams with higher Elo ratings generally tend to score more goals per match, as seen by the upward-sloping regression line. Since Elo summarizes long-term team performance and competitive strength before the tournament, we can see that this rating is a relevant predictor for offensive output. This positive association supports including Elo rating as a key feature in our model for predicting goals scored.

2) Goals vs Matches Played

Typically, teams that advance further in the tournament tend to play more matches. Because scoring requires opportunity, we wanted to study whether match count was related to goals scored.

```
plot_goals_vs_matches(wc) # match count from qualified_teams.csv
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

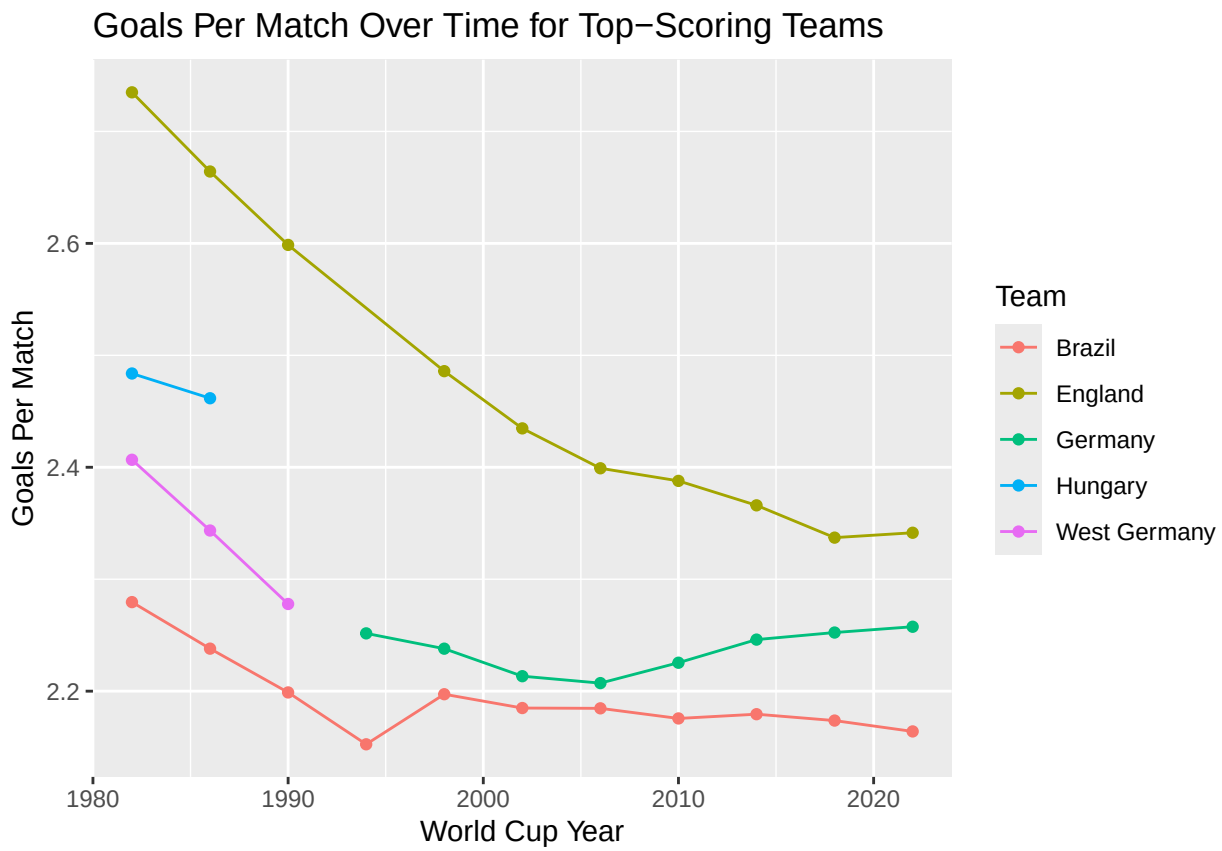


This plot compares how many matches a team plays in a World Cup tournament to the number of goals they score per match. Although there is a good amount of variation within each match count, the slight upward trend displayed by the red line suggests that teams playing more games (and advancing to later rounds) tended to receive higher scores, which logically does makes sense. Although the relationship is weaker than the one observed with Elo rating, the matches played still provides helpful information for modeling goal scoring.

3) Goals Over Time for Top Scoring Teams

To understand long-term scoring patterns, we decided to track the teams with the highest historical goals-per-match averages and plot their scoring across tournaments.

```
plot_goals_over_time(wc)
```

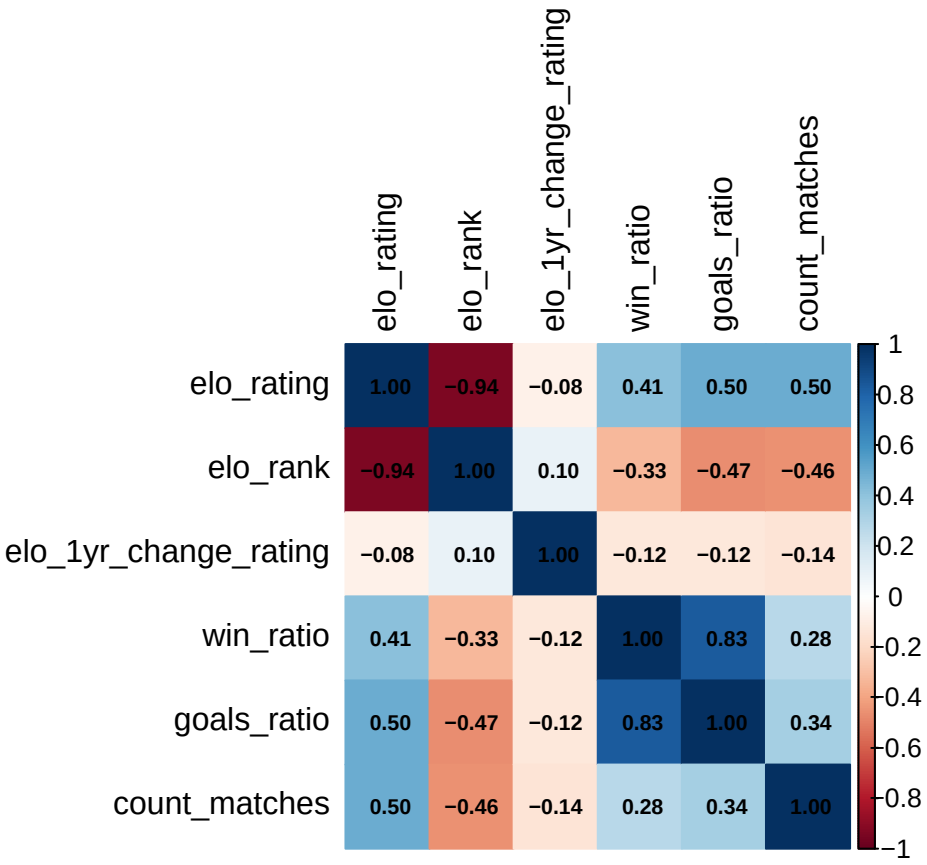


This line graph tracks how the highest-scoring teams in our dataset have performed across different World Cups. While the specific levels vary by team, the majority of them show a gradual decline or stabilization in goals per match over time. Ultimately, this reflects more broad trends in international soccer, including more balanced tournament competition and even stronger defensive structures. We can also see that historically strong teams like Brazil and Germany maintain relatively consistent scores. These patterns suggest that past scoring behavior can be useful for prediction and future scoring, but ultimately, team scoring ability evolves over time.

4) Correlation Heatmap

Here, we can visualize and assess how the key numeric variables in our dataset (Elo rating, recent performance, match count, and goal scoring) all relate with one another.

```
plot_corr_heatmap(wc)
```



As seen in this heatmap, goals per match shows strong positive correlation with win ratio, and moderate positive correlation with Elo rating. This indicates that the teams who win more and have higher pre-tournament strength tend to score more goals. Match count is also positively correlated to goals and Elo, showing that stronger teams have the ability to not only advance further, but also generate more scoring opportunities. Additionally, Elo rank is strongly negatively correlated with Elo rating as lower ranks indicate stronger teams. Overall, these relationships support using Elo-based and performance-based variables in our predictive model.