MarchMadness

```
library(readr)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
files <- list.files("March Madness Regular Season Data", pattern = "\\.csv$", full.names = TR


regular_season_df <- files |>
  lapply(function(f) read_csv(f, col_types = cols(.default = "c"))) |>
  bind_rows()
```

```
getwd()
```

```
[1] "/home/guest/sta345-final"
```

```
list.files()
```

```
[1] "March Madness Regular Season Data"
[2] "MarchMadness.pdf"
[3] "MarchMadness.qmd"
[4] "MarchMadness.rmarkdown"
[5] "sta345-final.Rproj"
[6] "tourney_results_mod_through_2025.csv"
```

```
list.files("March Madness Regular Season Data")
```

```
 [1] "ncaa_tr_regular_2000_2001.csv" "ncaa_tr_regular_2002_2003.csv"
 [3] "ncaa_tr_regular_2004_2005.csv" "ncaa_tr_regular_2006_2007.csv"
 [5] "ncaa_tr_regular_2008_2009.csv" "ncaa_tr_regular_2010_2011.csv"
 [7] "ncaa_tr_regular_2010_2025.csv" "ncaa_tr_regular_2012_2013.csv"
 [9] "ncaa_tr_regular_2014_2015.csv" "ncaa_tr_regular_2016_2017.csv"
[11] "ncaa_tr_regular_2018_2019.csv" "ncaa_tr_regular_2020_2021.csv"
[13] "ncaa_tr_regular_2022_2023.csv" "ncaa_tr_regular_2024_2025.csv"
```

```
tourney_result <- read_csv("tourney_results_mod_through_2025.csv")
```

```
Rows: 2585 Columns: 15
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): wloc, wteam_school, lteam_school, wteam_region, lteam_region
dbl (10): season, daynum, wteam, wscore, lteam, lscore, numot, wteam_seed, l...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
lower_better <- c(
  # Negative outcomes (your team)
  "Turnovers per Game",
  "Turnovers per Possession",
  "Turnovers per Offensive Play",
  "Personal Fouls per Game",
  "Personal Fouls per Possession",
  "Personal Fouls per Defensive Play",
```

```
  # Opponent stats - you want *less* of these
  "Opponent Points per Game",
  "Opponent Average Scoring Margin",
  "Opponent Floor %",
  "Opponent 1st Half Points per Game",
  "Opponent 2nd Half Points per Game",
  "Opponent Overtime Points per Game",
  "Opponent Points from 2 pointers",
  "Opponent Points from 3 pointers",
  "Opponent Shooting %",
  "Opponent Effective Field Goal %",
  "Opponent Three Point %",
  "Opponent Two Point %",
  "Opponent Free Throw %",
  "Opponent True Shooting %",
  "Opponent Field Goals Made per Game",
  "Opponent Three Pointers Made per Game",
  "Opponent Free Throws Made per Game",
  "Opponent Non-blocked 2 Pt %",
  "Opponent Offensive Rebounds per Game",
  "Opponent Offensive Rebounding %",
  "Opponent Assists per Game",
  "Opponent Assists per FGM",
  "Opponent Assists per Possession",
  "Opponent Assist / Turnover Ratio",
  "Opponent Win % - All Games",
  "Opponent Win % - Close Games",
  "Opponent Effective Possession Ratio",

  # Defensive efficiency (points allowed per 100 poss)
  "Defensive Efficiency",

  # Mixed/negative:
  # Opponent Defensive Rebounding % is bad for you
  "Opponent Defensive Rebounding %"
)
lower_better[!lower_better %in% names(regular_season_df)]
```

```
character(0)
```

```r
regular_season_unique <- regular_season_df |>
  distinct(school, year, .keep_all = TRUE)

stat_cols <- regular_season_unique |>
  select(-school, -year) |>
  names()
```

```r
team_season_ranks <- regular_season_unique |>
  group_by(year) |>
  mutate(
    # Stats where higher is better
    across(
      all_of(setdiff(stat_cols, lower_better)),
      ~ {
        x_num <- parse_number(as.character(.))
        rank(-x_num, ties.method = "average")
      }
    ),

    # Stats where lower is better
    across(
      all_of(lower_better),
      ~ {
        x_num <- parse_number(as.character(.))
        rank(x_num, ties.method = "average")
      }
    )
  ) |>
  ungroup()
```

```
Warning: There were 470 warnings in `mutate()`.
The first warning was:
i In argument: `across(...)`.
i In group 22: `year = "2021"`.
Caused by warning:
! 10 parsing failures.
row col expected actual
348  -- a number     --
349  -- a number     --
350  -- a number     --
351  -- a number     --
352  -- a number     --
```

```
... ... ........ ......
See problems(...) for more details.
i Run `dplyr::last_dplyr_warnings()` to see the 469 remaining warnings.
```

```
regular_season_unique %>%
  count(school, year) %>%
  filter(n > 1) %>%
  arrange(desc(n))
```

```
# A tibble: 0 x 3
# i 3 variables: school <chr>, year <chr>, n <int>
```

round of 64 games we are looking at

```
top_vs_bottom_round64 <- tourney_result |>
  mutate(
    low_seed  = pmin(wteam_seed, lteam_seed),
    high_seed = pmax(wteam_seed, lteam_seed)
  ) |>
  filter(
    round == 64,
    (low_seed == 1 & high_seed == 16) |
    (low_seed == 2 & high_seed == 15) |
    (low_seed == 3 & high_seed == 14) |
    (low_seed == 4 & high_seed == 13)
  )
```