Sports Analytics Final Project

2025-03-31

To summarize here is a list of cleaned data sets merged with roster data for 2024 games:

- `master_df` contains every Texas attack recorded in the season
- `by_match_df` contains hitting percentage for each player by match
- `season_df` contains hitting percentage for each player for the entire season

```
# import roster data
roster <- read_csv("~/Senior Year/Sports Analytics/Data/roster.csv")
```

```
## Rows: 20 Columns: 6
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (4): player_name, position, class, home_state
## dbl (2): player_number, height_inches
##
## 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.
```

```
# View(roster)
```

Texas Attack Data

The columns that will remain in the data set are as follows: player_name, player_number,input_type, output_type, position, match_date, skill, match_id, evaluation, attack_code, start_coordinate_x, end_coordinate_x, start_coordinate_y, end_coordinate_y, winning_attack, video_time, opponent, epv_in, epv_out, epv_added, height_inches, home_state, and class.

```
# Filter to only Texas Attack Data
ncaa_ut <- filter(ncaa, team=="University of Texas at Austin", skill=="Attack") |>
  # Select relevant columns prior to joining data to conserve computing power
  select(player_name, player_number,input_type, output_type, position, match_date, skill,
         match_id, evaluation, attack_code, start_coordinate_x, end_coordinate_x,
         start_coordinate_y, end_coordinate_y, winning_attack, video_time, opponent, epv_in, epv
_out, epv_added)
sec_ut <- filter(sec,team=="University of Texas at Austin", skill=="Attack") |>
  select(player_name, player_number,input_type, output_type, position, match_date, skill,
         match_id, evaluation, attack_code, start_coordinate_x, end_coordinate_x,
         start_coordinate_y, end_coordinate_y, winning_attack, video_time, opponent, epv_in, epv
_out, epv_added)
b12_ut <- filter(b12, team=="University of Texas at Austin", skill=="Attack") |>
  select(player_name, player_number,input_type, output_type, position, match_date, skill, match_
id, evaluation, attack_code, start_coordinate_x, end_coordinate_x,
         start_coordinate_y, end_coordinate_y, winning_attack, video_time, opponent, epv_in, epv
_out, epv_added)
```

```
# join datasets together add in player roster info
master_df <- full_join(ncaa_ut, sec_ut) |>
  full_join(b12_ut) |>
  mutate(player_name = case_when(player_name=="Player 10" ~ "Reagan Rutherford",
                                 .default = as.character(player_name))) |>
  inner_join(select(roster,player_name, height_inches,home_state),by='player_name')
```

```
## Joining with `by = join_by(player_name, player_number, input_type, output_type,
## position, match_date, skill, match_id, evaluation, attack_code,
## start_coordinate_x, end_coordinate_x, start_coordinate_y, end_coordinate_y,
## winning_attack, video_time, opponent, epv_in, epv_out, epv_added)`
## Joining with `by = join_by(player_name, player_number, input_type, output_type,
## position, match_date, skill, match_id, evaluation, attack_code,
## start_coordinate_x, end_coordinate_x, start_coordinate_y, end_coordinate_y,
## winning_attack, video_time, opponent, epv_in, epv_out, epv_added)`
```

Hitting Percentage of Players by Match

```
# hitting percentage of players by match
by_match_df <- master_df |>
  # select(player_name, player_number,match_id, opponent,evaluation, output_type,winning_attack,
epv_in, epv_out, epv_added) |>
  mutate(error = ifelse(evaluation=="Error", 1, 0)) |>
  group_by(player_name, match_id, opponent) |>
  mutate(attempts = n(),
         errors = sum(error, na.rm=T),
         kills = sum(winning_attack, na.rm=T),
         hitting_pctg = (kills - errors)/attempts) |>
  select(-winning_attack, -evaluation, -output_type) |>
  # Make sure attacker has at least 4 attempts to get a better idea of performance in match
  filter(attempts >=4)
```

Season Hitting Percentages by Player

```
# season hitting percentages with roster info
season_df <- master_df |>
  # select(player_name, player_number,match_id, opponent,evaluation, output_type,winning_attack,
epv_in, epv_out, epv_added) |>
  mutate(error = ifelse(evaluation=="Error", 1, 0)) |>
  group_by(player_name) |>
  mutate(attempts = n(),
           errors = sum(error, na.rm=T),
           kills = sum(winning_attack, na.rm=T),
           hitting_pctg = (kills - errors)/attempts) |>
  select(-winning_attack, -evaluation, -output_type) |>
  # Make sure attacker has at least 4 attempts to get a better idea of performance in match
  filter(attempts >=4)
```