Final Proposal STA478

Sakina Lord

2025-11-18

Introduction

The chosen dataset contains rugby match results and player data from the French Top 14 league. Every professional rugby match I have attended in person has been part of this league, which includes Union Bordeaux-Bègles, the team from my host city in France. Compared to more widely studied competitions such as HSBC Sevens or the Six Nations, the Top 14 is relatively underexplored. Personal interest also influenced this choice, as I played rugby union during the summer and it remains my favorite "ball sport."

This project uses two datasets: one containing 18 seasons of French Top 14 Rugby Union results, and another with per-player information.

- Top14 Results
- Player Information

Discussion of Variables

Top 14 The Top 14 dataset contains per-game data for 3,338 matches over 18 seasons involving 26 teams. In the raw dataset, all variables are character type. During cleaning, I converted *home_score* and *away_score* to numeric types. I also created:

- *delayed_game* — 1 if the match was postponed due to COVID-19, 0 otherwise
- *score_diff* — the score difference, positive if the home team won, negative if the away team won
- *winning_team* — indicates which team won

Additional key variables include *Season* (the two-year season span), *date* (match day), *home_team* and *away_team*, and *home_1-23* and *away_1-23* (the players for each team in that match).

Top 14 seasons are played over weekends denoted as J1 through J26, with each team having 13 home and 13 away games. The top six teams progress to the final phase, where 3rd-6th place teams compete in *Barrages*. The winner of the 3rd/6th place match faces the 2nd-ranked team, and the winner of the 4th/5th place match faces the 1st-ranked team in the *Semi*, both played at neutral venues. The *Finale* is contested between the winners of the semifinals at a neutral stadium, which may affect how the model interprets home and away designations. The *Access Match* pits the 13th-ranked team against the second division champion for a place in the next Top 14 season.

As shown in the first corrplot, score-related variables (*home_score*, *score_diff*, etc.) exhibit some correlation, whereas non-score variables show little to no correlation.

Players The Players dataset contains per-player information, including *birthdate*, *position*, and *team* for a given *competition* and *year*. It also tracks performance metrics such as *total_points_scored*, *matches_played*, *tries_scored*, and *matches_started*.

Data originates from itsrugby.fr, allowing me to make column names more descriptive. *player_id*, scraped from each player's URL, was cleaned to contain only the player name in uppercase, aligning it with the Top 14 dataset.

- Character variables: *player_id*, *birthdate*, *position*, *team*, *competition*
- Numeric variables: performance stats

Players data aggregates per-competition metrics, whereas the Top 14 dataset contains per-game statistics. This motivates a potential join between the two datasets. Before filtering for Top 14 matches, the dataset included 5,968 players across 26 years and 86 competitions, totaling 51,211 rows. Some entries contain erroneous data, such as a player listed in a 2097 competition (*player_id thomas-lainault-3896*).

Data Cleaning

A key decision is whether to use only Top 14 player data or include matches from other leagues. Filtering for Top 14 reduces the player dataset from 51,211 entries to 4,355, removing players from Irish, English, Welsh, and Scottish leagues such as the Six Nations. When converting scores to numeric type, I encountered "Delayed" and "TBD" values. TBD scores correspond to future matches and are removed. Delayed matches, often due to COVID-19, are retained to analyze pandemic impacts while ensuring score columns remain numeric.

A major challenge in merging the datasets is that many Top 14 players do not appear in the Players dataset. The raw datasets only share player names as identifiers, but in different formats. Top 14 contains over 4,000 unique players, while the cleaned Players dataset contains only 1,048, leaving individual statistics missing for 3,154 players. This gap exists across all seasons from 2005–2022, so truncating by season is not an option. Missing two-thirds of player data presents a substantial limitation if player stats are used for modeling.

Analysis

I am still unsure of what relationships to predict using this dataset, so an unsupervised approach using clustering could help reveal natural paterns in the data.I would also like to learn more about creating unsupervised models and how to interpret them. I'd like to use K-means clustering to identify groups within both player and match data. This might highlight which players or teams are most erratic or unpredictable, and could also be applied to a third dataset limited to only players with match records.

It may be also useful to predict what the probability of the home team winning is and to understand why. I could use a less interpretable random forest to just predict the final score of a matchup, and a logistic regression to understand what factors, such as whether the match is home or away, the presence of a player with a certain number of tries, or the day of the tournament, increase the score difference. I could also use the player data to assess whether aggression on the field, indicated by a yellow or red card is an asset or liability to a team.

Goals

In this project, I hope to complete at least 3 models: a KMeans clustering, a random forest which I might want to try to boost, and a well-crafted logistic model. I chose these models because they are relevant to my dataset (I don't have so many predictors that PCA would be effective) and because these are either new techniques or ones I am not fully comfortable with yet. I would also like to improve my ability to interpret,

penalize, and improve these models. I also hope to understand a little better how and, more importantly, when to apply penalties and choose tuning parameters when using real data.

I have also run into the issue of an incomplete dataset causing problems in modeling such as in my French capstone, and I wonder if some of the sampling techniques from early in the semester could help to describe distributions of some of this data. When joined, the resulting dataset is missing about 75% of player-match pairings, and I wonder if they have some distribution able to be modeled by a random sampling technique.

Finally, I want to work with a mildly challenging dataset. This dataset is not all numbers, has a lot of categorical data, requires joins, and is based, in part, on incomplete scraped data. When coerced into numeric formats, there are multiple binary variables; when there are multiple binary predictors I am unsure how to handle modelling, so I would like to get more comfortable with these types of datasets.

Code and Outputs

Load Data

```
top14 <- as.data.frame(read.csv("./data/top14_data.csv"))
players <- as.data.frame(read.csv("./data/itsrugby_players.csv"))

kable(str(top14, list.len = 12), label = "Top 14 Data Structure")
```

```
## 'data.frame':    3338 obs. of  54 variables:
##  $ season    : chr  "2005-2006" "2005-2006" "2005-2006" "2005-2006" ...
##  $ day       : chr  "J1" "J1" "J1" "J1" ...
##  $ date      : chr  "2005-08-20" "2005-08-20" "2005-08-20" "2005-08-20" ...
##  $ home_team : chr  "Aviron Bayonnais" "Section Paloise" "USA Perpignan" "RC Toulon" ...
##  $ away_team : chr  "Stade Toulousain" "ASM Clermont" "SU Agen" "Biarritz Olympique PB" ...
##  $ home_score: chr  "12" "16" "34" "10" ...
##  $ away_score: chr  "26" "28" "9" "20" ...
##  $ stadium   : chr  "" "" "" "" ...
##  $ home_1    : chr  "JEAN MARIE USANDISAGA" "STÉPHANE DELPUECH" "PERRY FRESHWATER" "MICHEL PERIE" ..
##  $ home_2    : chr  "CHRISTOPHE LAURENT" "FERNANDO GUATIERI" "MICHEL KONIECKIEWICZ" "CAMILLE TRAVERSA
##  $ home_3    : chr  "JÉRÉMY TOMULI" "OLIVIER SOURGENS" "SÉBASTIEN BOZZI" "EUSEBIO GUINAZU" ...
##  $ home_4    : chr  "CÉDRIC BERGEZ" "KARL RUDZKI" "COLIN GASTON" "NICOLAAS SMIT" ...
##   [list output truncated]
```

```
kable(str(players), label = "Player Data Structure")
```

```
## 'data.frame':    51211 obs. of  16 variables:
##  $ player_id  : chr  "esteban-abadie-4289" "esteban-abadie-4289" "esteban-abadie-4289" "esteban-abadi
##  $ birthdate  : chr  "01/12/1997" "01/12/1997" "01/12/1997" "01/12/1997" ...
##  $ pos        : chr  "Third Row" "Third Row" "Third Row" "Third Row" ...
##  $ year       : int  2022 2021 2021 2020 2020 2019 2019 2017 2022 2021 ...
##  $ team       : chr  "Brive" "Brive" "Brive" "Brive" ...
##  $ competition: chr  "Top 14" "European Rugby Challenge Cup" "Top 14" "European Rugby Challenge Cup"
##  $ pts        : num  0 0 10 0 0 0 0 0 0 0 ...
##  $ played     : num  2 1 20 2 8 2 2 1 1 2 ...
##  $ start      : num  2 0 14 2 3 2 2 0 1 1 ...
##  $ try        : num  0 0 2 0 0 0 0 0 0 0 ...
##  $ pen        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ dp         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tr         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ yellow     : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ red        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ min        : num  160 29 1185 136 315 ...
```

Data Cleaning

```
# PLAYER DATA CLEANING
# Rename Player columns to be more descriptive and filter for only Top14 players
players_clean <- players %>% rename(position = pos,
```

```
                            total_points_scored = pts,
                            matches_played = played,
                            matches_started = start,
                            tries_scored = try,
                            penalty_goals_scored = pen,
                            drop_goals_scored = dp,
                            conversions = tr,
                            yellow_cards = yellow,
                            red_cards = red,
                            total_minutes_played = min) %>% filter(competition == "Top 14")
# Reformat player_id to just have text and make an ID to join the data on
players_clean$player_name <- players_clean$player_id %>%
  str_replace("-\\d+$", "") %>%   # remove the trailing -1234
  str_replace_all("-", " ") %>%   # replace hyphens with spaces
  str_to_upper()%>%               # convert to upper case
  stri_trans_general("Latin-ASCII") # Get rid of all the accents and weird letters

str(players_clean)
```

```
## 'data.frame':    4355 obs. of  17 variables:
##  $ player_id           : chr  "esteban-abadie-4289" "esteban-abadie-4289" "esteban-abadie-4289" "este
##  $ birthdate           : chr  "01/12/1997" "01/12/1997" "01/12/1997" "01/12/1997" ...
##  $ position            : chr  "Third Row" "Third Row" "Third Row" "Third Row" ...
##  $ year                : int  2022 2021 2020 2019 2017 2022 2021 2020 2019 2018 ...
##  $ team                : chr  "Brive" "Brive" "Brive" "Brive" ...
##  $ competition         : chr  "Top 14" "Top 14" "Top 14" "Top 14" ...
##  $ total_points_scored : num  0 10 0 0 0 0 0 52 10 0 ...
##  $ matches_played      : num  2 20 8 2 1 1 15 19 13 19 ...
##  $ matches_started     : num  2 14 3 2 0 1 9 18 10 17 ...
##  $ tries_scored        : num  0 2 0 0 0 0 0 4 2 0 ...
##  $ penalty_goals_scored: num  0 0 0 0 0 0 0 10 0 0 ...
##  $ drop_goals_scored   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ conversions         : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ yellow_cards        : num  0 0 1 0 0 0 1 0 0 1 ...
##  $ red_cards           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_minutes_played: num  160 1185 315 139 24 ...
##  $ player_name         : chr  "ESTEBAN ABADIE" "ESTEBAN ABADIE" "ESTEBAN ABADIE" "ESTEBAN ABADIE" ..
```

```
#TOP 14 DATA CLEANING
# Convert scores to Numeric data type, factor days, covert date to date datatype,
# and add winner and point difference columns to Top14 data
top14_clean <- top14 %>% filter(away_score != "TBD") %>%
  filter(home_score != "TBD") %>%
  mutate(delayed_game = ifelse((away_score == "Delayed")|(home_score == "Delayed"), 1, 0 )) %>%
  mutate(home_score = ifelse(away_score == "Delayed", 0, home_score)) %>%
  mutate(away_score = ifelse(away_score == "Delayed", 0, away_score)) %>%
  mutate(home_score = as.numeric(home_score)) %>%
  mutate(away_score = as.numeric(away_score)) %>%
  mutate(score_diff = home_score - away_score) %>%
  mutate(winning_team = ifelse(score_diff > 0, home_team,
                        ifelse(score_diff == 0, "Null", away_team)))
str(top14_clean)
```

```
## 'data.frame':    3289 obs. of  57 variables:
##  $ season      : chr  "2005-2006" "2005-2006" "2005-2006" "2005-2006" ...
##  $ day         : chr  "J1" "J1" "J1" "J1" ...
##  $ date        : chr  "2005-08-20" "2005-08-20" "2005-08-20" "2005-08-20" ...
##  $ home_team   : chr  "Aviron Bayonnais" "Section Paloise" "USA Perpignan" "RC Toulon" ...
##  $ away_team   : chr  "Stade Toulousain" "ASM Clermont" "SU Agen" "Biarritz Olympique PB" ...
##  $ home_score  : num  12 16 34 10 16 15 26 29 52 29 ...
##  $ away_score  : num  26 28 9 20 34 10 20 8 9 23 ...
##  $ stadium     : chr  "" "" "" "" ...
##  $ home_1      : chr  "JEAN MARIE USANDISAGA" "STÉPHANE DELPUECH" "PERRY FRESHWATER" "MICHEL PERIE"
##  $ home_2      : chr  "CHRISTOPHE LAURENT" "FERNANDO GUATIERI" "MICHEL KONIECKIEWICZ" "CAMILLE TRAVE
##  $ home_3      : chr  "JÉRÉMY TOMULI" "OLIVIER SOURGENS" "SÉBASTIEN BOZZI" "EUSEBIO GUINAZU" ...
##  $ home_4      : chr  "CÉDRIC BERGEZ" "KARL RUDZKI" "COLIN GASTON" "NICOLAAS SMIT" ...
##  $ home_5      : chr  "SIULÉO LAFAIALI'I" "GARRICK MORGAN" "NATHAN HINES" "FRANCK ALAZET" ...
##  $ home_6      : chr  "GUILLAUME COMBES" "PIERRE SOM" "GREGORY LE CORVEC" "GREGORY LABADZE" ...
##  $ home_7      : chr  "PHIL DAVIES" "ALEXANDRU MANTA" "GUILLAUME BORTOLASO" "MOHAMMED DRIDI" ...
##  $ home_8      : chr  "LOUIS MASSABEAU" "PATRICK TABACCO" "OVIDIU TONITA" "HAROLD KARELE" ...
##  $ home_9      : chr  "GREGORY SUDRE" "CHRISTOPHE LAUSSUCQ" "NICOLAS DURAND" "LUDOVIC LOUSTAU" ...
##  $ home_10     : chr  "GERARD FRASER" "GONZALO QUESADA" "NICOLAS LAHARRAGUE" "MARTIN VICKERS-PEARSON
##  $ home_11     : chr  "BENJAMIN LHANDE" "BRENDON DANIEL" "MATTHIEU BOURRET" "RACHID OURAK" ...
##  $ home_12     : chr  "SÉBASTIEN ROQUE" "JEAN-EMMANUEL CASSIN" "DAVID MARTY" "FRANCK COMBA" ...
##  $ home_13     : chr  "JAMES MAC LAREN" "JEAN CHARLES CISTTACQ" "JEAN PHILIPPE GRANDCLAUDE" "DAVID D
##  $ home_14     : chr  "BENJAMIN THIERY" "MATHIEU DOURTHE" "PASCAL BOMATI" "GRÉGORY TUTARD" ...
##  $ home_15     : chr  "RICHARD DOURTHE" "JEAN-MARC SOUVERBIE" "JULIEN LAHARRAGUE" "PATRICE TEISSEIRE
##  $ home_16     : chr  "EDUARD COETZEE" "DAVID LAPERNE" "MARIUS TINCU" "PHILIP FITZGERALD" ...
##  $ home_17     : chr  "GRANT HILL" "JOHANE LURO" "AGUSTIN LOPRESTTI" "ROMAIN PAIRAULT" ...
##  $ home_18     : chr  "JORGE GARCIA" "PATRICK FURET" "CHRISTOPHE PORCU" "SOANE TOEVALU" ...
##  $ home_19     : chr  "MIKAERA TEWHATA" "VINCENT FORGUES" "SAMUELI DAWAI NAULU" "SAREL JACOBUS LOUW"
##  $ home_20     : chr  "VIATCHESLAV GRATCHEV" "FABIEN CIBRAY" "BRUNO ROLLAND" "BENOÎT MARFAING" ...
##  $ home_21     : chr  "MATHIEU SIRO" "LIONEL BEAUXIS" "SEBASTIEN DESCONS" "FREDERIC ARNIAUD" ...
##  $ home_22     : chr  "HENRI VERMIS" "GREGORY PUYO" "DIEGO GIANNANTONIO" "AUBIN HUEBER" ...
##  $ home_23     : chr  "GREGORY MENKARSKA" "LAURENT EMMANUELLI" "JEAN JACQUES CRENCA" "KASIANO LEALAM
##  $ away_1      : chr  "YANNICK BRU" "BRICE MIGUEL" "JALIL NARJISSI" "BENOIT AUGUST" ...
##  $ away_2      : chr  "OMAR HASAN" "MARTIN SCELZO" "PATRICK BLANCO" "BENOÎT LECOULS" ...
##  $ away_3      : chr  "TREVOR BRENNAN" "ALEXANDRE AUDEBERT" "DAMIEN FEVRE" "JEROME THION" ...
##  $ away_4      : chr  "ROMAIN MILLO CHLUSKI" "GONZALO LONGO" "KIRILL KULEMIN" "DAVID COUZINET" ...
##  $ away_5      : chr  "JEAN BOUILHOU" "MICHEL DIEUDE" "GAVIDI RATUVA" "SERGE BETSEN" ...
##  $ away_6      : chr  "GREGORY LAMBOLEY" "SAM BROOMHALL" "AARON PERSICO" "IMANOL HARINORDOQUY" ...
##  $ away_7      : chr  "FINAU MAKA" "RAPHAËL CHANAL" "MATTHIEU LIEVREMONT" "THOMAS LIÈVREMONT" ...
##  $ away_8      : chr  "JEAN BAPTISTE ELISSALDE" "PIERRE MIGNONI" "NICOLAS MORLAES" "JULIEN DUPUY" ..
##  $ away_9      : chr  "JEAN FREDERIC DUBOIS" "JEAN-BAPTISTE DAMBIELLE" "FRANCOIS GELEZ" "JULIEN PEYR
##  $ away_10     : chr  "CEDRIC HEYMANS" "JULIEN MALZIEU" "RUPENI CAUCAUNIBUCA" "DIMITRI YACHVILI" ...
##  $ away_11     : chr  "YANNICK JAUZION" "TONY MARSH" "SYLVAIN MIRANDE" "DENIS LISON" ...
##  $ away_12     : chr  "FLORIAN FRITZ" "GONZALO CANALE" "CORNELIUS STOLTZ" "THIBAULT DUVALLET" ...
##  $ away_13     : chr  "VINCENT CLERC" "AURELIEN ROUGERIE" "LUC LAFFORGUE" "ARAMBURU FEDERICO MARTIN"
##  $ away_14     : chr  "CLEMENT POITRENAUD" "ANTHONY FLOCH" "PEPITO ELHORGA" "DAMIEN TRAILLE" ...
##  $ away_15     : chr  "JULIAN FIORINI" "MARIO LEDESMA AROCENA" "CHRISTIAN CALIFANO" "SERELI BOBO" ..
##  $ away_16     : chr  "WILLIAM SERVAT" "RÉMI VAQUIN" "ALESSIO GALASSO" "MARC BAGET RABAROU" ...
##  $ away_17     : chr  "ISITOLO MAKA" "LOÏC JACQUET" "JEAN-MICHEL PARENT" "OLIVIER OLIBEAU" ...
##  $ away_18     : chr  "JULIEN LE DEVEDEC" "DAVID BARRIER" "FABRICE CULINE" "PETRU BALAN" ...
##  $ away_19     : chr  "ANTOINE BATTUT" "EMMANUEL ETIEN" "FRANÇOIS TANDONNET" "JIMMY MARLU" ...
##  $ away_20     : chr  "XAVIER GARBAJOSA" "JAMIE CUDMORE" "JEROME MIQUEL" "DENIS AVRIL" ...
##  $ away_21     : chr  "FREDERIC MICHALAK" "DAVIT ZIRAKASHVILI" "JEAN-BAPTISTE RUE" "GUILLAUME BERGOS
##  $ away_22     : chr  "" "" "" "" ...
```

```
##  $ away_23     : chr  "" "" "" "" ...
##  $ delayed_game: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ score_diff  : num  -14 -12 25 -10 -18 5 6 21 43 6 ...
##  $ winning_team: chr  "Stade Toulousain" "ASM Clermont" "USA Perpignan" "Biarritz Olympique PB" ...
```

Joining Datasets

```
# Create a new dataframe with stats per-player
# length(unique(players_clean$player_name))

# Pivot Top 14 data so that each player has their own row
  # Code and regex matching from ChatGPT
top14_clean_long <- top14_clean %>%
  pivot_longer(
    # match the name + position number
    cols = matches("^(home|away)_[0-9]+$"),
    names_to = c("side", "position"),
    names_pattern = "(home|away)_(\\d+)",
    values_to = "player_name"
  ) %>%
  mutate(position = as.integer(position)) %>%
  mutate(year = as.integer(str_sub(season, 1, 4)))

# Remove all the weird accents, etc. in player names
  # line from ChatGPT
top14_clean_long$player_name <- stri_trans_general(top14_clean_long$player_name, "Latin-ASCII")

str(top14_clean_long)
```

```
## tibble [151,294 x 15] (S3: tbl_df/tbl/data.frame)
##  $ season      : chr [1:151294] "2005-2006" "2005-2006" "2005-2006" "2005-2006" ...
##  $ day         : chr [1:151294] "J1" "J1" "J1" "J1" ...
##  $ date        : chr [1:151294] "2005-08-20" "2005-08-20" "2005-08-20" "2005-08-20" ...
##  $ home_team   : chr [1:151294] "Aviron Bayonnais" "Aviron Bayonnais" "Aviron Bayonnais" "Aviron Bay
##  $ away_team   : chr [1:151294] "Stade Toulousain" "Stade Toulousain" "Stade Toulousain" "Stade Toul
##  $ home_score  : num [1:151294] 12 12 12 12 12 12 12 12 12 12 ...
##  $ away_score  : num [1:151294] 26 26 26 26 26 26 26 26 26 26 ...
##  $ stadium     : chr [1:151294] "" "" "" "" ...
##  $ delayed_game: num [1:151294] 0 0 0 0 0 0 0 0 0 0 ...
##  $ score_diff  : num [1:151294] -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 ...
##  $ winning_team: chr [1:151294] "Stade Toulousain" "Stade Toulousain" "Stade Toulousain" "Stade Toul
##  $ side        : chr [1:151294] "home" "home" "home" "home" ...
##  $ position    : int [1:151294] 1 2 3 4 5 6 7 8 9 10 ...
##  $ player_name : chr [1:151294] "JEAN MARIE USANDISAGA" "CHRISTOPHE LAURENT" "JEREMY TOMULI" "CEDRIC
##  $ year        : int [1:151294] 2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
```

```
# Join with player data on year and player name
joined <- top14_clean_long %>%
  left_join(players_clean, by = c("player_name", "year"))
str(joined)
```

```
## tibble [151,731 x 30] (S3: tbl_df/tbl/data.frame)
```

```
##  $ season              : chr [1:151731] "2005-2006" "2005-2006" "2005-2006" "2005-2006" ...
##  $ day                 : chr [1:151731] "J1" "J1" "J1" "J1" ...
##  $ date                : chr [1:151731] "2005-08-20" "2005-08-20" "2005-08-20" "2005-08-20" ...
##  $ home_team           : chr [1:151731] "Aviron Bayonnais" "Aviron Bayonnais" "Aviron Bayonnais" "Av
##  $ away_team           : chr [1:151731] "Stade Toulousain" "Stade Toulousain" "Stade Toulousain" "Sta
##  $ home_score          : num [1:151731] 12 12 12 12 12 12 12 12 12 12 ...
##  $ away_score          : num [1:151731] 26 26 26 26 26 26 26 26 26 26 ...
##  $ stadium             : chr [1:151731] "" "" "" "" ...
##  $ delayed_game        : num [1:151731] 0 0 0 0 0 0 0 0 0 0 ...
##  $ score_diff          : num [1:151731] -14 -14 -14 -14 -14 -14 -14 -14 -14 -14 ...
##  $ winning_team        : chr [1:151731] "Stade Toulousain" "Stade Toulousain" "Stade Toulousain" "Sta
##  $ side                : chr [1:151731] "home" "home" "home" "home" ...
##  $ position.x          : int [1:151731] 1 2 3 4 5 6 7 8 9 10 ...
##  $ player_name         : chr [1:151731] "JEAN MARIE USANDISAGA" "CHRISTOPHE LAURENT" "JEREMY TOMULI"
##  $ year                : int [1:151731] 2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...
##  $ player_id           : chr [1:151731] NA NA NA NA ...
##  $ birthdate           : chr [1:151731] NA NA NA NA ...
##  $ position.y          : chr [1:151731] NA NA NA NA ...
##  $ team                : chr [1:151731] NA NA NA NA ...
##  $ competition         : chr [1:151731] NA NA NA NA ...
##  $ total_points_scored : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ matches_played      : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ matches_started     : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ tries_scored        : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ penalty_goals_scored: num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ drop_goals_scored   : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ conversions         : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ yellow_cards        : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ red_cards           : num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
##  $ total_minutes_played: num [1:151731] NA NA NA NA NA NA NA NA NA NA ...
```

Missing Player Data

```
missing_players <- joined %>%
  filter(is.na(player_id)) %>%
  select(player_name, year)

sprintf("Missing data on %d players", length(unique(missing_players$player_name)))
```

```
## [1] "Missing data on 3154 players"
```

```
# One weird outlier
x <- players %>% filter(competition == "Yves du Manoir Cup")
str(x)
```

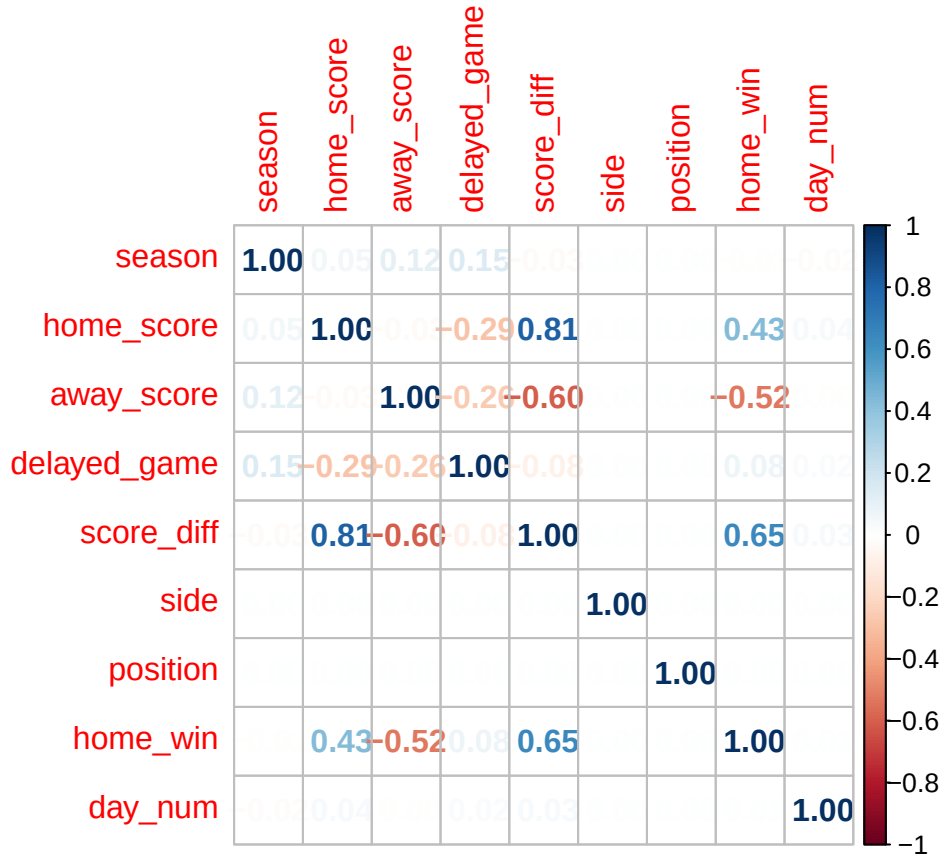
```
## 'data.frame':    1 obs. of  16 variables:
##  $ player_id  : chr "thomas-lainault-3896"
##  $ birthdate  : chr "28/12/1993"
##  $ pos        : chr "Lock"
##  $ year       : int 2097
##  $ team       : chr "Rouen"
```

```
##  $ competition: chr "Yves du Manoir Cup"
##  $ pts        : num NA
##  $ played     : num NA
##  $ start      : num NA
##  $ try        : num NA
##  $ pen        : num NA
##  $ dp         : num NA
##  $ tr         : num NA
##  $ yellow     : num NA
##  $ red        : num NA
##  $ min        : num NA
```

Correlations for Modeling

```
top14_numerical_only <- top14_clean_long %>% mutate(season = as.integer(str_sub(season, 1, 4))) %>%
  select(-year) %>% mutate(side = ifelse(side == "home", 1, 0)) %>%
  mutate(home_win = ifelse(score_diff < 0, 0, 1)) %>%
  mutate(day = factor(day)) %>% mutate(day_num = as.numeric(day)) %>%
  select(-home_team, -away_team, -stadium, -player_name, -winning_team, -date, -day)
```

```
corrplot(cor(top14_numerical_only), method = "number")
```



What if we just looked at one particular player (UBB's Scrum Half)

```
select.player <- "MAXIME LUCU"

top14_one_player <- top14_clean_long %>% filter(player_name == select.player)%>%
  mutate(season = as.integer(str_sub(season, 1, 4))) %>%
  mutate(side = ifelse(side == "home", 1, 0)) %>%
  mutate(home_win = ifelse(score_diff < 0, 0, 1)) %>%
  mutate(day = factor(day)) %>% mutate(day_num = as.numeric(day)) %>%
  mutate(win = ifelse(side == home_win, 1, 0)) %>%
  select(-home_team, -away_team, -stadium, -player_name,
         -winning_team, -date, -day, -year)
```

```
corrplot(cor(top14_one_player), method = "number")
```

```
## Warning in cor(top14_one_player): the standard deviation is zero
```

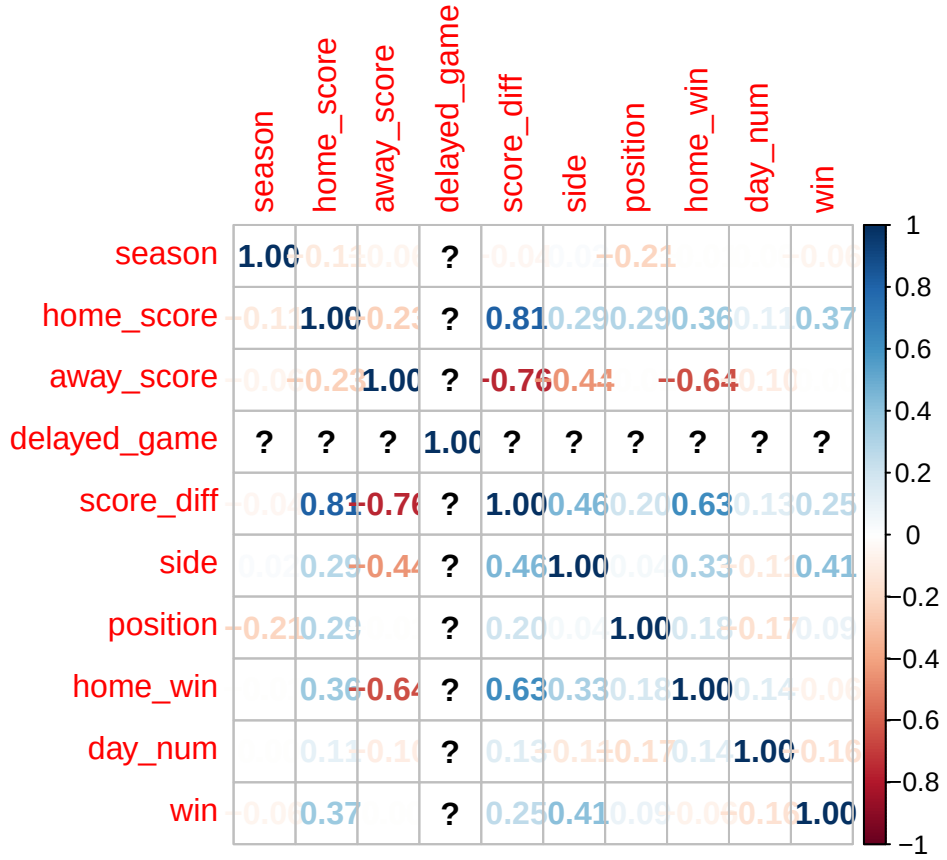


Figure 1: One Player (Maxime Lucu) Variable Corrplot

```
# ChatGPT used for the factor recode
players_numeric <- players_clean %>%
  mutate(position = recode(position,
                           "Third Row" = 6,
                           "Scrum half" = 9,
                           "Wing" = 11,
```

```
                            "Hooker" = 3,
                            "Centre" = 12,
                            "Prop" = 1,
                            "Lock" = 4,
                            "Full back" = 15,
                            "Fly Half" = 10,
                            .default =  0))%>%
  mutate(position = as.integer(position)) %>%
  mutate(birth_year = as.integer(str_sub(birthdate, 7, 11)))%>%
  mutate(team = as.integer(factor(team))) %>%
  select(-player_id, -birthdate, -competition, - player_name)
#str(players_numeric)
```

```
# Not sure why birth year has question marks
corrplot(cor(players_numeric))
```

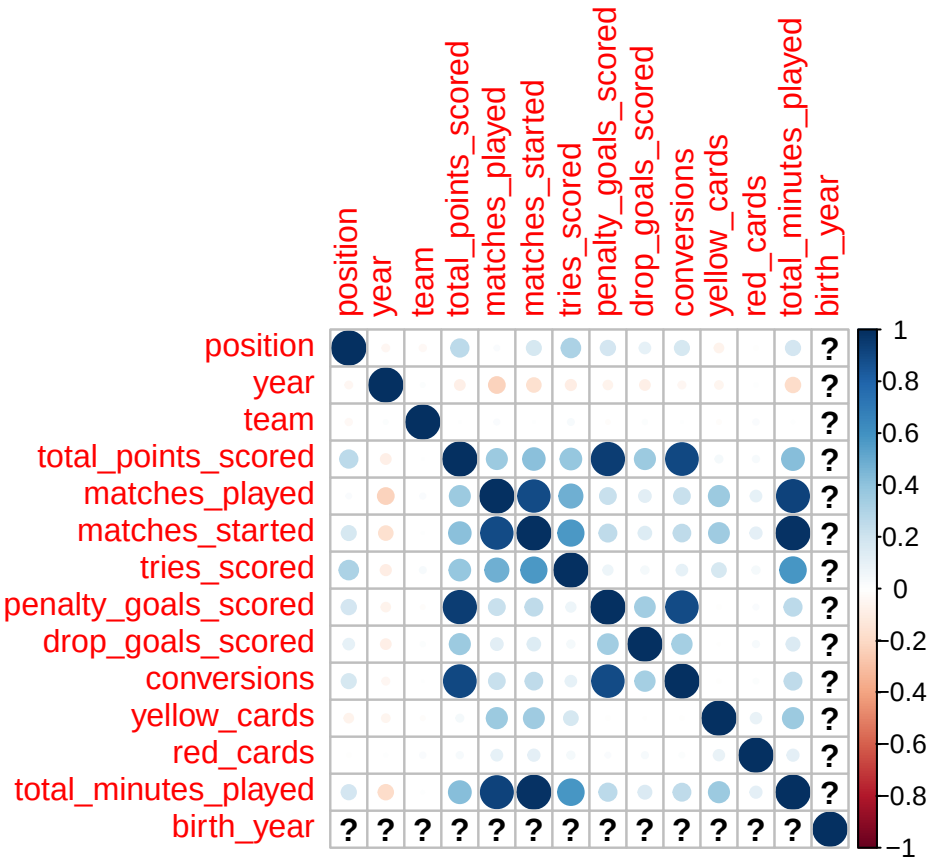


Figure 2: Players Numerical Variable Corrplot