~~~
---
title: "Data Analysis Project"
author: "Caroline Koutsos, Khushboo Rathore"
date: "28 April 2023"
  html_document:
    theme: cerulean
    highlight: pygments
    toc: true
    toc_float:
      collapsed: true
      smooth_scroll: false
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## Load libraries

Loading required libraries for this analysis.

```{r echo=FALSE, message=FALSE}
library(tidyverse)
library(lubridate)
library(janitor)
library(dplyr)
library(ggplot2)
library(corrr)
library(ggrepel)
library(stringr)
```

## Load and Cleaning Data

```{r}
### LOAD AND CLEAN DATA ###

wsoccer_matchstat_2020 <- read_csv("data/ncaa_womens_soccer_matchstats_2020.csv", guess_max = 5000)
wsoccer_matchstat_2021 <- read_csv("data/ncaa_womens_soccer_matchstats_2021.csv", guess_max = 5000)
wsoccer_matchstat_2022 <- read_csv("data/ncaa_womens_soccer_matchstats_2022.csv", guess_max = 5000)

wsoccer_playerstat_2020 <- read_csv("data/ncaa_womens_soccer_playerstats_2020.csv", guess_max = 5000)
wsoccer_playerstat_2021 <- read_csv("data/ncaa_womens_soccer_playerstats_2021.csv", guess_max = 5000)
wsoccer_playerstat_2022 <- read_csv("data/ncaa_womens_soccer_playerstats_2022.csv", guess_max = 5000)

wsoccer_teams_2020 <- read_csv("url_csvs/NCAA Women's Soccer - 2020.csv") %>%
  clean_names()
wsoccer_teams_2021 <- read_csv("url_csvs/NCAA Women's Soccer - 2021.csv") %>%
  clean_names()
wsoccer_teams_2022 <- read_csv("url_csvs/NCAA Women's Soccer - 2022.csv") %>%
  clean_names()

team_links_2020 <- read_csv("url_csvs/ncaa_womens_soccer_teamurls_2020.csv") %>%
  clean_names()
team_links_2021 <- read_csv("url_csvs/ncaa_womens_soccer_teamurls_2021.csv") %>%
  clean_names()
team_links_2022 <- read_csv("url_csvs/ncaa_womens_soccer_teamurls_2022.csv") %>%
  clean_names()

# Create a column for the team number taken from the link
team_links <- team_links_2020 %>%
  rbind(team_links_2021, team_links_2022) %>%
  distinct(school, .keep_all = TRUE) %>%
  mutate(team_id = str_extract(playerstatsurl, "(?<=team/).*?(?=/stats)"))

# Find the column names, create a key
# col_names <- colnames(wsoccer_matchstat_2020)
# Data Dictionary: https://docs.google.com/spreadsheets/d/1lnEauTJ62LcCfWJdwIHUitD6Hai_-1YCDVsEbwiTSG8/edit#gid=0

# Clean 2020 match data so that we can rbind()
match_20 <- wsoccer_matchstat_2020 %>%
  mutate(season = 2020) %>%
  select(-points, -na.x, -na.y, -defensive_points)

# Clean 2021 match data so that we can rbind()
match_21 <- wsoccer_matchstat_2021 %>%
  mutate(season = 2021) %>%
  select(-gwg, -defensive_gwg)

# Clean 2022 match data so that we can rbind()
match_22 <- wsoccer_matchstat_2022 %>%
  mutate(season = 2022) %>%
  select(-gwg, -defensive_gwg)

# All matches across all three seasons
match_all <- match_20 %>%
  rbind(match_21, match_22) %>%
  mutate(ha_number = case_when(
    home_away == "Home"~ 0,
    home_away == "Away"~ 1,
    TRUE ~ NA
  )) %>%
  mutate(team = case_when(
    team == "FDU Knights"~"Fairleigh Dickinson Knights",
    team == "Kansas City Roos"~"Kansas City Kangaroos",
    team == "Nebraska Huskers"~"Nebraska Cornhuskers",
    team == "Houston Baptist Huskies"~"Houston Christian Huskies",
~~~

```
    TRUE~team
  ))

# Clean 2021 team data so that we can rbind()
teams_21 <- wsoccer_teams_2021 %>%
  select(-so_g, -gwg, -games) %>%
  mutate(season = 2021)

# Clean 2022 team data so that we can rbind()
teams_22 <- wsoccer_teams_2022 %>%
  select(-so_g, -gwg, -games) %>%
  mutate(season = 2022)

# Clean 2020 team data so that we can rbind()
# cols_20 <- names(wsoccer_teams_2020)

teams_20 <- wsoccer_teams_2020 %>%
  select(url, institution, conference, goals, assists, -points, sh_att, fouls, red_cards, yellow_cards, gc, goal_app, ggs, 
goalie_min_plyd, ga, saves, shutouts, combined_sho, g_wins, g_loss, d_saves, corners, pk, pk_att) %>%
  mutate(season = 2020)

# Clean 2020 player data so that we can rbind()
players_20 <- wsoccer_playerstat_2020 %>%
  select(season, team, jersey, full_name, roster_name, first_name, last_name, yr, pos, gp, gs, goals, assists, points, sh_att, fouls, 
red_cards, yellow_cards, gc, goal_app, ggs, goalie_min_plyd, ga, gaa, saves, sv_pct, combined_sho, shutouts, g_wins, g_loss, d_saves, 
corners, pk, pk_att, gwg) %>%
  mutate(season = 2020)

# Clean 2021 player data so that we can rbind()
players_21 <- wsoccer_playerstat_2021 %>%
  select(-so_g, -games) %>%
  mutate(season = 2021)

# Clean 2022 player data so that we can rbind()
players_22 <- wsoccer_playerstat_2022 %>%
  select(-so_g, -games) %>%
  mutate(season = 2022)

# All players across three seasons
players_all <- players_20 %>%
  rbind(players_21, players_22) %>%
  mutate(team = case_when(
    team == "FDU Knights"~"Fairleigh Dickinson Knights",
    team == "Kansas City Roos"~"Kansas City Kangaroos",
    team == "Nebraska Huskers"~"Nebraska Cornhuskers",
    team == "Houston Baptist Huskies"~"Houston Christian Huskies",
    TRUE~team
  ))

# All teams across three seasons
all_teams <- teams_20 %>%
 rbind(teams_21, teams_22) %>%
  mutate(institution = case_when(
    institution == "Albany (NY)"~"UAlbany",
    institution == "Dixie St."~"Utah Tech",
    TRUE~institution
  ))

# Regular expression of universities
regex_unis <- regex(paste0("\\b", paste(unique(all_teams$institution), collapse = "\\b|\\b"), "\\b"))

# Crosswalk of team name to mascot
mascot_crosswalk <- players_all %>%
  distinct(team) %>%
  mutate(mascot = str_squish(sub(regex_unis, "", team)))

# Crosswalk of team name, institution, mascot, conference
team_crosswalk <- all_teams %>%
  mutate(team_id = str_extract(url, "(?<=team/).*?(?=/)")) %>%
  distinct(institution, conference, season, .keep_all = TRUE) %>%
  left_join(team_links, by = "team_id") %>%
  left_join(mascot_crosswalk, by = c("school" = "team")) %>%
  select(institution, mascot, school, conference, team_id) %>%
  rename(team = school) %>%
    mutate(mascot = case_when(
    institution == "VMI"~"Cadets",
    TRUE~mascot
  )) %>%
  filter(!is.na(mascot))

# Previous crosswalk with seasons
tc_season <- team_crosswalk %>%
  inner_join(all_teams, by = c("institution", "conference")) %>%
  distinct(institution, conference, season, .keep_all = TRUE) %>%
  select(institution:team_id, season)
```

Question One:

General comparison of performance at Home vs Away games:
  Do teams get more shutouts at Home games?
  Do players get carded more at Away games?
  Does being at a foreign field impact wins and losses aka W/L ratios?

Answer:

Our analysis seems to indicate that teams have a significant advantage at home games. There was a significant difference in win percent, shutouts, and yellow cards for each team between home and away games. However, there was not a significant difference in the number of fouls usually given in home vs away games. It would be interesting to continue this by looking at difference in referees and if there tend to be stricter refs in certain conferences or at certain universities which could contribute to the cards. It would also be interesting to look at home and away with the additional context of a different field surface (actual grass vs turf, etc.)

In noticing that teams have a significant advantage at home, does this contribute to the importance of the “win on the road” mentality? Also, does this very among different sports or does it remain the same across sports, collegiate and professional? with the number of fouls given not being significant, does the “fouls” include only yellow and red cards or anytime the player gets a foul without a card? Also, it would be interesting to see if the number of set pieces contributes to wins as well, because there is a phrase in women’s soccer: “games are won or lost on setpieces”. If we were to do this again, maybe we could see the amount of free kicks, corner kicks, and pks that led to a goal or win.

Overall, further analysis of this would probably be the most newsworthy. While the difference in overall performance isn't very new, understanding the reasons behind these changes could help teams and change the understanding of the sport.

``` {r}

## We'll be talking on Tuesday about correlations - which share some similarities with t-tests - and I think those could be useful for you in determining how important those home field advantages are. You also could do some multiple regression to see if a combination of factors add up to a meaningful relationship (https://www.thescoop.org/sports/multipleregression.html).

matches_wld <- match_all %>%
  # Group games by home/away and win-loss-draw for each team
  group_by(team, home_away, outcome) %>%
  count() %>%
  pivot_wider(names_from = outcome, values_from = n) %>%
  replace(is.na(.), 0) %>%
  # Create a column that has the Win-Loss-Draw ratio for the teams across all seasons
  mutate(wld_ratio = paste0(Win, "-", Loss, "-",Draw)) %>%
  # Create a column that shows the win percentage for home vs away games across all seasons
  mutate(win_percent = Win/(Win+Loss+Draw)*100)

matches_compares <- match_all %>%
  # Assign a numeric value to each outcome
  mutate(outcome = case_when(
    str_detect(outcome,"Loss")~0,
    str_detect(outcome,"Win")~1,
    str_detect(outcome,"Draw")~.5,
    TRUE~NA
  )) %>%
  group_by(team, home_away, ha_number) %>%
  # Create summary statistics for a variety of
  summarize(
    w_l_d = mean(outcome),
    shutouts = sum(shutouts),
    all_red = sum(red_cards),
    all_yellow = sum(yellow_cards),
    foul_mean = mean(fouls),
    foul_median = median(fouls),
    score_mean = mean(team_score),
    score_med = median(team_score),
    goal_rate = sum(goals)/sum(sh_att),
  ) %>%
  left_join(matches_wld, by = c("team", "home_away"))

# Boxplots that show wins, yellow cards, and mean of fouls
boxplot(matches_compares$win_percent ~ matches_compares$home_away)
boxplot(matches_compares$all_yellow ~ matches_compares$home_away)
boxplot(matches_compares$foul_mean ~ matches_compares$home_away)

# Separate dataframes into data by home and away games
teams_home <- matches_compares %>%
  filter(home_away == "Home")

teams_away <- matches_compares %>%
  filter(home_away == "Away")

# T-Test to see home and away differences.
t.test(teams_home$win_percent, teams_away$win_percent, paired=TRUE)
t.test(teams_home$shutouts, teams_away$shutouts, paired=TRUE)
t.test(teams_home$all_yellow, teams_away$all_yellow, paired=TRUE)
t.test(teams_home$foul_mean, teams_away$foul_mean, paired=TRUE)

# Look at correlations for some variables, including home/away, win percent, goal rate, mean fouls, and mean score
matches_compares %>%
  select(ha_number, win_percent, goal_rate, foul_mean, score_mean) %>%
  correlate()

# Look at a linear regression for win percent by home/away, goal rate and score mean.
summary(lm(matches_compares$win_percent ~ matches_compares$ha_number + matches_compares$goal_rate + matches_compares$score_mean))

```
Question Two Original:

Do more popular numbers (attacking players) (10, 11, 9, 7) equate to more goals scored per season?

We wanted to pivot our question 2 because after a lengthy and fruitful discussion, we realized that I was basing my "popular numbers in soccer" on the men's side of soccer, particularly professional. To be honest, culturally, the men's and women's side of soccer are slightly different. So based on (my) experience in women's soccer since I was 3 years old, I would understand how the more popular numbers would be different depending on gender. Also, more men watch professional men's soccer (I'm assuming) - Caroline

```{r}
# # Specifically collect statistics for any players who have jersey numbers 10, 11, 9, 7, and 23.
# player_filter <- players_all %>%
#   filter(jersey == 10| jersey == 11| jersey == 9| jersey == 7| jersey == 23) %>%
#   select(-full_name, -points, -red_cards:-corners)
#
# # Look at the statistics for players outside of those jersey numbers
# player_anti <- players_all %>%
#   filter(!(jersey == 10| jersey == 11| jersey == 9| jersey == 7| jersey == 23)) %>%
```

~~~
#   select(-full_name, -points, -red_cards:-corners)

# # Get some statistics for each jersey number we like
# jersey_players <- player_filter %>%
#   # Asked Chat GPT: "How do I replace the NA values in every numeric column in a R dataframe with 0?"
#   mutate_if(is.numeric, ~replace(., is.na(.), 0)) %>%
#   group_by(jersey) %>%
#   summarize(
#     mean_gp = mean(gp),
#     mean_goals = mean(goals),
#     median_goals = median(goals),
#     goal_shtrat = sum(goals)/sum(sh_att)*100,
#     mean_fouls = mean(fouls),
#     pk_success = sum(pk)/sum(pk_att)*100,
#     gwg = mean(gwg)
#   )
#
# # Get statistics for all other jersey numbers
# other_players <- player_anti %>%
#   mutate_if(is.numeric, ~replace(., is.na(.), 0)) %>%
#   group_by(jersey) %>%
#   summarize(
#     mean_gp = mean(gp),
#     mean_goals = mean(goals),
#     median_goals = median(goals),
#     goal_shtrat = sum(goals)/sum(sh_att)*100,
#     mean_fouls = mean(fouls),
#     pk_success = sum(pk)/sum(pk_att)*100,
#     gwg = mean(gwg)
#   )
```
~~~

Question Two Revised:

Is there a significant relationship in performance and jersey number? Why are certain numbers more popular and which numbers have the best performance? Why do players choose the numbers that they do?

Answer:

If we define success or positive performance in women's college soccer as goals scored and the amount of game winning goals scored, we can look at the means for both of those statistics and see that the number 10 is first for popularity, mean_goals and mean_gwg. According to https://yoursoccerhome.com/the-number-10-in-soccer-why-its-so-significant/ , the number 10 is typically worn by an attacking midfielder/forward. Out of 380 players in the pop_to_names dataframe, there are only 41 players labeled as defenders wearing the number 10. That means the majority of players wearing #10 are forwards and midfielders. According to yoursoccerhome, starting in England in 1928, numbers were used to identify players on the field, and the numbers 1-11 were used. Each position had a number. For example, 1 was a goalkeeper, 2-5 was the backline, 6 8 and 10 were the midfielders, and 7, 9, and 11 were attacking players. Nowadays, those numbers aren't limited to positions, but the association still exists and coaches will reinforce those position numbers by teaching their players the positions using those numbers. They will use those numbers during film, while teaching tactics etc. Though the origins come from identifying certain players, the number 10 became less of a method of identification and more of a recognition of a certain player as time went on. The role of number 10 was to be a "playmaker." Players such as Messi, Diego Maradona, Carli Lloyd, Marta and Pele are all well known footballers that wore/wear these numbers. In women's soccer, Carli Lloyd and Marta were iconic playmakers who were creative, tactically gifted and continuously put the ball in the back of the net.

~~~
```{r}

# Your work here left me wondering what the distribution is like for all numbers - in other words, whether your choice of popular numbers (23, for example, doesn't seem to be massively popular compared to some others). I'd encourage you to start broad and then narrow down your focus based on your findings and theories.

# Dataframe of all players, only once per their jersey number
player_dist <- players_all %>%
  distinct(roster_name, jersey, .keep_all = TRUE)

gk_players <- player_dist %>%
  filter(str_detect(pos, "GK"))

# Exclude goalkeepers since their statistics seem to skew data a bit
nongk_player_by_jersey <- player_dist %>%
  anti_join(gk_players) %>%
  group_by(jersey) %>%
  summarize(
    num_players = n(),
    mean_goals = mean(goals, na.rm = TRUE),
    median_goals = median(goals, na.rm = TRUE),
    mean_gwg = mean(gwg, na.rm = TRUE),
    mean_shtrat = mean((goals/sh_att)*100, na.rm = TRUE),
    mean_gp = mean(gp, na.rm = TRUE),
    mean_fouls = mean(fouls, na.rm = TRUE)
  )

# Create a chart that shows popular numbers and mean goals
jersey_chart <- nongk_player_by_jersey %>%
  ggplot(aes(x = num_players, y = mean_goals)) +
  geom_point(size = 1, shape = 1, color = "black") +
  labs(
    x = "Number of Players",
    y = "Mean Goals"
  )

jersey_chart

# Jersey numbers with over 300 distinct players
over_threehund <- nongk_player_by_jersey %>%
  filter(num_players >= 300)

# Players who use the jersey numbers mentioned above
pop_to_names <- over_threehund %>%
~~~

```
  inner_join(player_dist, by = "jersey")
```

Question 3:

Upperclassmen vs Underclassmen: what years typically dominate the starting lineup in different conferences

Answer:

Though we find that the teams who start a higher percentage of seniors that are on their roster have higher average goals scored, we have an interesting caveat to this. 197 teams out of the total 373 start more freshmen out of the freshmen that they have (52%). This was derived from the amount of rows in the top_under dataframe. This means that though teams who rely heavily on their senior class tend to have more goals, a majority of teams will use their freshmen. This is notable because out of one of the top conferences for women's soccer, the ACC, 7 of 14 teams play majority freshmen and sophomore, which is half of those teams in that conference. We think that this is significant because of the recent news in US women's soccer, where more youth players are getting professional contracts. The NWSL just started an "under-18" program, and Chloe Ricketts, a 15 year old, just became the youngest individual to join the NWSL under a professional contract. She's playing for the Washington Spirit. Just before that, Alyssa Thompson made history for being the first high school student to be drafted to the NWSL, and was drafted as No.1 overall pick to Angel City FC. She was also just called up to the USWNT senior roster for the 2023 Women's World Cup in Australia. Olivia Moultrie (17) and Jaedyn Shaw (18) are also major breakthrough young players who are expected to make some noise in the professional league.

We think this is interesting because it shows how rising youth talent is on the rise currently in women's soccer. We think it's interesting that as all of this is happening with young players entering the draft and being noticed/fostered by NWSL clubs, more college level teams are utilizing their younger players in their starting lineups from what they have in their rosters. Though these don't necessarily correlate, we think that it would also be worth testing to see how many high school seniors graduate a semester early to go play for their college teams in the spring before their freshman year. We say this because UMD women's soccer had 3 freshmen that graduated high school early to come to play with the team in the spring. If we had the data on how many high school seniors did this, maybe we could find a trend that points to this idea of younger players accelerating into higher levels of play faster than they used to. If we had this, we could do a t test or something to see how the number of hs seniors graduating early is correlated to a higher number of starting freshmen, and if it's statistically significant. Follow up question if we were to continue: does this "acceleration" contribute to more college underclassmen playing for their respective national teams?

Things to consider for this question though are that if teams changed conferences, the dataframe top_under includes them in both conferences which could be considered an issue. However, it accounts for a possible change in percent of which grade was played after they changed conferences.
https://www.goal.com/en-us/lists/us-girls-youth-soccer-system-uswnt-stars/blt3599216dee5320a7
https://www.cbsnews.com/colorado/news/chloe-ricketts-soccer-nwls-washington-spirit-15-year-old-player/
https://www.cnn.com/2023/01/14/football/alyssa-thompson-us-soccer-profile-spt-intl/index.html

```{r}

games_per_seasonteam <- match_all %>%
  group_by(season, team) %>%
  count()

team_unique <- all_teams %>%
  group_by(institution, conference) %>%
  summarize(
    avg_goals = round(mean(goals),2),
    avg_shutouts = round(mean(shutouts),2),
    avg_fouls = round(mean(fouls),2)
  ) %>%
  inner_join(tc_season, by = c("institution", "conference")) %>%
  distinct(institution, conference, season, .keep_all = TRUE)

player_team <- players_all %>%
  select(season:gs, team, -pos, -full_name) %>%
  left_join(games_per_seasonteam, by = c("season", "team")) %>%
  left_join(team_unique, by = c("team", "season")) %>%
  rename(season_games = n)

player_by_year <- player_team %>%
  group_by(roster_name, institution, yr, conference, season) %>%
  summarize(
    percent_start = round(gs/season_games*100,2)
  ) %>%
  mutate(percent_start = case_when(
    is.na(percent_start)~0,
    TRUE~percent_start
  ))

institution_year <- player_team %>%
  group_by(roster_name, institution, yr, conference, season) %>%
  summarize(
    percent_start = round(gs/season_games*100,2)
  ) %>%
  mutate(percent_start = case_when(
    is.na(percent_start)~0,
    TRUE~percent_start
  )) %>%
  group_by(institution, conference, yr) %>%
  summarize(
    starter = mean(percent_start)
  ) %>%
  pivot_wider(names_from = "yr", values_from = "starter") %>%
  select(institution, conference, Fr, So, Jr, Sr) %>%
  left_join(team_unique, by = c("institution", "conference"))

start_play <- player_team %>%
  group_by(roster_name, institution, yr, conference, season) %>%
  summarize(
    start_play = round(gs/gp*100,2)
  ) %>%
  mutate(start_play = case_when(
    is.na(start_play)~0,
    TRUE~start_play
  )) %>%
  group_by(institution, conference, yr) %>%
  summarize(
    starter = mean(start_play)
  ) %>%
  pivot_wider(names_from = "yr", values_from = "starter") %>%
  select(institution, conference, Fr, So, Jr, Sr) %>%
  left_join(team_unique, by = c("institution", "conference")) %>%
  distinct(institution, conference, .keep_all = TRUE)

top_under <- start_play %>%
  filter(So > 50 | Fr > 50)

conference_stats <- institution_year %>%
  group_by(conference) %>%
  summarize(
    fr_se = round(mean(Fr),2),
    so_se = round(mean(So),2),
    jr_se = round(mean(Jr),2),
    sr_se = round(mean(Sr),2),
    mean_goals = round(mean(avg_goals),2),
    mean_foul = round(mean(avg_fouls),2)
  )

conference_startplay <- start_play %>%
  group_by(conference) %>%
  summarize(
    fr_sp = round(mean(Fr),2),
    so_sp = round(mean(So),2),
    jr_sp = round(mean(Jr),2),
    sr_sp = round(mean(Sr),2),
    mean_goals = round(mean(avg_goals),2),
    mean_foul = round(mean(avg_fouls),2)
  )

comp_eligible_played <- conference_stats %>%
  left_join(conference_startplay, by = c("conference", "mean_goals", "mean_foul")) %>%
  mutate(fr_dif = fr_sp - fr_se) %>%
  mutate(so_dif = so_sp - so_se) %>%
  mutate(jr_dif = jr_sp - jr_se) %>%
  mutate(sr_dif = sr_sp - sr_se) %>%
  select(conference, contains("dif"), contains("mean"))

```

Question 4:

Team with the most shutouts vs team with most goals scored: which team is more successful on avg? Does the strength of an offense or defense impact match result more?

Answer:

Obviously, teams that have a high mean shutouts and a high mean goals have a higher win percentage. Once we start to take a closer look and separate the teams out into multiple categories based on their shutouts and goals, we start to see a distinction between the impact of shutouts and goals on winning.

We looked specifically at Power 5 Schools, which is made up of schools that are in either the ACC, SEC, Big 10, Big 12 or Pac-12. In our graphic, we plotted the means of the average goals and average shutouts across the power 5. When we look at teams who perform above average on shutouts and below on goals as compared to teams who perform below average on shutouts and above on goals, we can see that the second case tends to have a higher win percent.

Similarly, when looking at all teams (keeping in mind that teams in multiple conferences over the three years have multiple points), we see a similar trend. Teams needed to get shutouts for over half their matches to have a comparable win percent to teams scoring two goals a match.

```{r}
team_wld <- match_all %>%
  # Group games by home/away and win-loss-draw for each team
  group_by(team, outcome, season) %>%
  count() %>%
  pivot_wider(names_from = outcome, values_from = n) %>%
  replace(is.na(.), 0) %>%
  left_join(tc_season, by = c("team", "season")) %>%
  clean_names()

matches_od <- match_all %>%
  left_join(tc_season, by = c("team", "season")) %>%
  left_join(team_wld, by = c("team", "season", "conference", "institution", "mascot")) %>%
  group_by(team, conference) %>%
  summarize(
    tot_shut = sum(shutouts),
    avg_shut = mean(shutouts),
    tot_goals = sum(goals),
    avg_goals = mean(goals),
    win_percent = round(sum(win)/(sum(win)+sum(loss)+sum(draw))*100, 2)
  )

match_avg <- matches_od %>%
  ungroup() %>%
  summarise(
    goal_avg = mean(avg_goals),
    shut_avg = mean(avg_shut)
    )

top_shut <- matches_od %>%
  arrange(desc(avg_shut)) %>%
  head(100) %>%
  mutate(status = "shutouts")
```

```
top_goals <- matches_od %>%
  arrange(desc(avg_goals)) %>%
  head(100) %>%
  mutate(status = "goals")

top_shut_nogoals <- matches_od %>%
  anti_join(top_goals) %>%
  arrange(desc(avg_shut)) %>%
  head(100) %>%
  mutate(status = "shutouts_not_goals")

top_goal_noshut <- matches_od %>%
  anti_join(top_shut) %>%
  arrange(desc(avg_goals)) %>%
  head(100) %>%
  mutate(status = "goals_not_shutouts")

pfive_avg <- matches_od %>%
  filter(str_detect(conference, "ACC|SEC|Big Ten|Big 12|Pac-12")) %>%
  mutate(pfive = "yes") %>%
  group_by(pfive) %>%
  summarise(
    goal_avg = mean(avg_goals),
    shut_avg = mean(avg_shut)
    )

stats_pfive <- matches_od %>%
  filter(str_detect(conference, "ACC|SEC|Big Ten|Big 12|Pac-12"))

pfive_chart <- stats_pfive %>%
  ggplot +
    geom_point(
      aes(x = avg_goals, y = avg_shut, size = win_percent, color = conference),
      alpha = .5
      ) +
  # geom_point(
  #   data=inner_join(stats_pfive, top_shut_nogoals),
  #   aes(x = avg_goals, y = avg_shut, size = win_percent),
  #   color="red",
  #   alpha = .5
  #   ) +
  # geom_point(
  #   data=inner_join(stats_pfive, top_goal_noshut),
  #   aes(x = avg_goals, y = avg_shut, size = win_percent, color),
  #   color="blue",
  #   alpha = .5
  #   ) +
  scale_size(range = c(1, 8), name="Win Percent") +
  geom_vline(xintercept = pfive_avg$goal_avg) +
  geom_hline(yintercept = pfive_avg$shut_avg) +
  labs(
    title = "In Power 5 women's soccer, goals are more important than shutouts",
    x = "Mean Goals",
    y = "Mean Shutouts"
  )

pfive_chart

### DID NOT SHOW SUPER RELEVANT INFO COMPARED TO ALL_TEAMS ###

# not_pfive <- matches_od %>%
#   filter(!str_detect(conference, "ACC|SEC|Big Ten|Big 12|Pac-12")) %>%
#   mutate(pfive = "no") %>%
#   group_by(pfive) %>%
#   summarise(
#     goal_avg = mean(avg_goals),
#     shut_avg = mean(avg_shut)
#     )
#
# stats_npfive <- matches_od %>%
#   filter(!str_detect(conference, "ACC|SEC|Big Ten|Big 12|Pac-12"))
#
#
# npfive_chart <- stats_npfive %>%
#   ggplot +
#     geom_point(
#       aes(x = avg_goals, y = avg_shut, size = win_percent, color = conference),
#       alpha = .5
#       ) +
#   # geom_point(
#   #   data=inner_join(stats_pfive, top_shut_nogoals),
#   #   aes(x = avg_goals, y = avg_shut, size = win_percent),
#   #   color="red",
#   #   alpha = .5
#   #   ) +
#   # geom_point(
#   #   data=inner_join(stats_pfive, top_goal_noshut),
#   #   aes(x = avg_goals, y = avg_shut, size = win_percent, color),
#   #   color="blue",
#   #   alpha = .5
#   #   ) +
#   scale_size(range = c(1, 8), name="Win Percent") +
#   geom_vline(xintercept = not_pfive$goal_avg) +
#   geom_hline(yintercept = not_pfive$shut_avg) +
#   labs(
#     x = "Mean Goals",
#     y = "Mean Shutouts"
```

~~~
#   )
#
# npfive_chart

all_team_wstatus <- matches_od %>%
  left_join(top_shut_nogoals) %>%
  left_join(top_goal_noshut, by = c("team", "conference", "tot_shut", "avg_shut", "avg_goals", "tot_goals", "win_percent")) %>%
  mutate(status = case_when(
    !is.na(status.x) & !is.na(status.y)~"Avg Goals, Avg Shutouts",
    !is.na(status.x)~"Low Goals, High Shutouts",
    !is.na(status.y)~"High Goals, Low Shutouts",
    avg_goals > 1.5~"High Goals, High Shutouts",
    TRUE~"Low Goals, Low Shutouts"
  )) %>%
  select(-status.y, -status.x)

all_team_chart <- all_team_wstatus %>%
  ggplot +
    geom_point(
      aes(x = avg_goals, y = avg_shut, size = win_percent, color = status),
      alpha = .7
      ) +
  # geom_point(
  #   data=top_shut_nogoals,
  #   aes(x = avg_goals, y = avg_shut, size = win_percent),
  #   color= "lightblue",
  #   alpha = .9
  #   ) +
  # geom_point(
  #   data=top_goal_noshut,
  #   aes(x = avg_goals, y = avg_shut, size = win_percent),
  #   color= "orange",
  #   alpha = .5
  #   ) +
  scale_size(range = c(1, 8), name="Win Percent") +
  geom_vline(xintercept = match_avg$goal_avg) +
  geom_hline(yintercept = match_avg$shut_avg) +
  labs(
    title = "Across Division I women's soccer teams, win percent is more related to mean goals",
    x = "Mean Goals",
    y = "Mean Shutouts"
  )


all_team_chart
```
~~~

Question 5:

Success rate of different types of mascots: (animals, mythical creatures, humans, etc.)?

Answer:

ANIMAL: (golden, purple, screaming) eagles, (runnin', lady) bulldogs, wildcats, falcons, zips, (lady) hornets, (lady) tigers, (lady) panthers, (lady) lions, razorbacks, cougars, cardinals, bruins, broncos, terriers, bison(s), bulls, roadrunners, mustangs, lancers, camels, mocs, chanticleers, buffaloes, rams, bluejays, (blue) hens, hornets, stags, owls, gators, hoya*, lopes (antelopes), Rainbow Wahine, (christian) huskies, (lady) jaguars, bengals, redbirds, Hoosiers, gaels, dolphins,   Gamecocks, JMU dukes, kangaroos, Roos, sharks, golden flashes, LMU (CA) Lions, leopards, greyhounds, Red Foxes, Thundering Herd, Terrapins (terps), golden gophers, (golden) grizzlies, lobos, aggies, tar heels, ospreys, bobcats, sooners, monarchs, mavericks, ducks, beavers, nittany lions, Mastodons, spiders, broncs, Peacocks, jackrabbits, coyotes, Salukis, Bonnies, tommies, horned frogs, longhorns, anteaters, retrievers, utes, beacons, catamounts, hokies, leathernecks, badgers, great danes, penguins

MYTHICAL: (sun, blue) devils, titans, (golden) griffins, (blue, lady) demons, demon deacons, dragons, phoenix, devilettes, Billikens, blazers, tritons

PEOPLE: governors, (lady) braves, mountaineers, (black, FDU, Scarlet) knights, matadors, chippewas, buccaneers, 49ers, vikings, (blue, red) raiders, flyers, pioneers, duquesne dukes*, (lady) pirates, islanders, colonels, seminoles, paladins, colonials, crusaders, vandals, explorers, trojans, Lady Techsters, jaspers, minutewomen, cowgirls, warriors, spartans, racers, highlanders, midshipmen, cornhuskers, huskers, lumberjacks, norse, fighting irish, (lady) rebels, quakers, pilots, friars, boilermakers, royals, Ladyjacks, Aztecs, Toreros, Dons, Saints, Hatters, Texas, Volunteers, cadets, gauchos, miners, vaqueros, trailblazers, VMI cadets, commodores, cavaliers, musketeers

NATURE: crimson tide, cyclones, Sycamores, (lady) flame, beach, hurricanes, buckeyes, waves, red storm, orange, golden hurricane, flames

MISC: big red, big green, purple aces, crimson, pride, Fighting Illini, Ragin' Cajuns, Ramblers, mean green, blue hose, the Red Flash, rockets, lady topper, tribe

For the mascot list, it seems that when we look on it purely based on mascot, it looks like mythical teams seem to be the best overall in stats (besides shutouts). Nature teams seem to be fouled the most, and they're pretty across the board for all of them. Miscellaneous stats are slightly random because of the size of it. In mean goals, mythical comes out ahead slightly. If we actually look at how they compare to each other across all matches, (we look at win percent in reverse) teams wiht mascots of miscellaneous, nature and people have a 1/3 of winning against mythical teams. Animal has a 43% chance of winning against mythical, which is higher, but mythical usually wins out. Overall though, the majority of the matchups are pretty even, (40 or 50%) so it's not that much of a significance. If you're making a team, we'd recommend to go with a mythical mascot (just in case you go off of pure luck). We would want to see how certain animals or people do against each other (for example, eagles vs eagles and see whether certain matchups in the same category are significant).

Caveat: the groups vary in size, so if a bunch of the animal teams (for example) do really poorly, it will reflect poorly on the whole group.

~~~
```{r}

# If you're going to go down this road, then you absolutely should answer the matchup question (how well to animals do vs people, etc).
Same for identical mascot name matchups - how often do they occur? You should get really creative here.

team_sums <- all_teams %>%
  group_by(institution, conference) %>%
  summarize(
    tot_goals = sum(goals),
~~~

```
    tot_shutouts = sum(shutouts),
    tot_fouls = sum(fouls)
  ) %>%
  distinct(institution, .keep_all = TRUE)

regex_peoplemascot <-
regex("governors|braves|mountaineers|duke|knights|matadors|chippewas|buccaneers|49ers|vikings|raiders|flyers|pioneers|pirates|islanders|co

mascot_list <- match_all %>%
  left_join(tc_season, by = c("team", "season")) %>%
  group_by(team, mascot, institution) %>%
  summarize(
    tot_yc = sum(yellow_cards)
  ) %>%
  inner_join(team_sums, by = "institution") %>%
  mutate(mas_category = case_when(
    str_detect(str_to_lower(mascot), "devil|titan|griffin|demon|dragon|phoenix|billiken|blazer|triton")~"mythical",
    str_detect(str_to_lower(mascot), regex_peoplemascot)~"people",
    str_detect(str_to_lower(institution), "vmi")~"people",
    str_detect(str_to_lower(mascot), "tide|cyclones|sycamores|flame|beach|hurricane|buckeyes|waves|storm|orange")~"nature",
    str_detect(str_to_lower(mascot), "big red|big green|purple aces|crimson|pride|illini|cajuns|ramblers|mean green|blue hose|red
flash|rockets|lady topper|tribe")~"miscellaneous",
    TRUE~"animal"
  )) %>%
  replace(is.na(.), 0)

mascot_list_alt <- mascot_list %>%
  select(mascot, mas_category)

mascot_matches <- match_all %>%
  select(team, team_id, opponent, opponent_id, season, outcome) %>%
  mutate(team_id = as.character(team_id)) %>%
  inner_join(tc_season, by = c("team_id", "season")) %>%
  select(team.x, team_id, opponent, opponent_id, mascot, season, outcome) %>%
  rename(mascot_team = mascot, team = team.x) %>%
  mutate(opponent_id = as.character(opponent_id)) %>%
  inner_join(tc_season, by = c("opponent_id" = "team_id", "season")) %>%
  select(team.x:mascot, -institution) %>%
  rename(mascot_opponent = mascot, team = team.x) %>%
  inner_join(mascot_list_alt, by = c("mascot_team" = "mascot")) %>%
  rename(category_team = mas_category, team = team.x) %>%
  select(-team.y) %>%
  left_join(mascot_list_alt, by = c("mascot_opponent" = "mascot")) %>%
  rename(category_opponent = mas_category, team = team.x) %>%
  select(-team.y) %>%
  distinct(team_id, opponent_id, season, .keep_all = TRUE)


mascot_comp <- mascot_matches %>%
  group_by(category_team, category_opponent, outcome) %>%
  summarize(
    count = n()
  ) %>%
  pivot_wider(names_from = outcome, values_from = count) %>%
  clean_names() %>%
  mutate(win_percent = round(win/(win+loss+draw)*100,2))

by_mascot <- mascot_list %>%
  group_by(mas_category) %>%
  summarize(
    mean_fouls = mean(tot_fouls),
    med_fouls = median(tot_fouls),
    mean_goals = mean(tot_goals),
    med_goals = median(tot_goals),
    med_shut = median(tot_shutouts)
  )

boxplot(mascot_list$tot_goals ~ mascot_list$mas_category)
boxplot(mascot_list$tot_shutouts ~ mascot_list$mas_category)

```

Overall Reflections

Overall, there were a lot more complications than anticipated in this project. We dealt with a lot of different chaos in terms of the teams not being standardized and having to find unique ways of connecting dataframes. For future steps on question 1, we would want to specifically talk to some teams and possibly track them through a couple of home and away games in a play-by-play style. What seems to be making the players uncomfortable or causes them to perform differently?

Caveats and Limitations:

- Some statistics in 2020 are not present in 2021 and 2022, and vice versa
- We do not have data about the teams as a whole (i.e. what division they are in, what conference they are in)
- 2020 had only some conferences playing matches since it was during COVID, also fewer players so some may have redshirted, opted-out, etc.
- Matches are not classified at playoff games or normal games, so we would need to check the dates.
- No game count or shots on goal data for 2020 players, not sure if this was a scrape problem or something else.
- Data in Question 4 was divided by conference and team. If teams changed conferences, especially when moving up to a harder one, they have two datapoints in the chart, and one is often extremely small.
- Due to various lacks of standardization, some answers may not be completely accurate. We tried our best to standardize everything, but it is possible that we missed something.
- Some team data may have been lost while trying to join datasets that were not super friendly with each other