STAT 184 Course Project — Final

Benjamin Ofori-Kuragu

Introduction

In this project, I'm exploring how college football win rates vary across different conferences and what underlying statistical factors contribute and go along with a program's long-term success. Using my primary dataset from Kaggle which contains historical win rates and team records I'll compare conferences like the SEC and Big Ten while accounting for the recent re-alignment changes. My two secondary datasets, which contain detailed offensive and defensive statistics from the 2013 and 2023 seasons, helps me see things over a longer period of time and how things can change over a decade span. This will allow me to analyze specific performance metrics that contribute to success. Together, these datasets will help me better understand both conference-level trends and program-level dominance over time.

Primary Data Set

Where the Data Came From:
The primary dataset comes from Kaggle, titled ***College Football Dataset***. The link to the dataset can be found here at: https://www.kaggle.com/datasets/cvergnolle/football-5. It was put together from various public sources, including NCAA reports and public sports databases.

Relevance to the Questions:

This dataset directly tracks win-loss records, conferences, and seasons, allowing me to calculate win rates for teams across different conferences over multiple years. It provides the foundation to compare win rates between SEC and Big Ten teams and to investigate how realignment has shifted conference strength.

Imported Data Inspection:

```
library(tidyverse)
cfb_primary <- read_csv("College Football.csv")
glimpse(cfb_primary)
```

```
Rows: 107
Columns: 22
$ Rank                   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
$ School                 <chr> "Texas", "Georgia", "Michigan", "Oklahoma", "Al~
$ State                  <chr> "Texas", "Georgia", "Michigan", "Oklahoma", "Al~
$ City                   <chr> "Austin", "Athens", "Ann Arbor", "Norman", "Tus~
$ `Win Rate`             <chr> "64%", "89%", "79%", "79%", "90%", "90%", "54%"~
$ Conference             <chr> "Big 12", "SEC", "Big Ten", "Big 12", "SEC", "B~
$ `Football Revenue`     <chr> "$162,450,371", "$155,951,449", "$131,403,197",~
$ `Football Profit`      <chr> "$110,899,705", "$94,229,326", "$78,994,662", "~
$ `Football Expenses`    <chr> "$51,550,666", "$61,722,123", "$52,408,535", "$~
$ `Undergrad Enrollment` <dbl> 41309, 30714, 32695, 21294, 32458, 46123, 25379~
$ `Football Attendence`  <dbl> 91939, 92746, 108763, 77795, 98720, 96756, 8614~
$ `Stadium Size`         <dbl> 100119, 92746, 107601, 82112, 101821, 104944, 8~
$ `% Fraternity`         <dbl> 0.12, 0.22, 0.08, 0.30, 0.27, 0.06, 0.24, 0.17,~
$ `% Sorority`           <dbl> 0.16, 0.33, 0.17, 0.37, 0.40, 0.10, 0.45, 0.20,~
$ `Basketball Revenue`   <chr> "$23,958,575", "$12,029,584", "$21,141,253", "$~
$ `Basketball Profit`    <chr> "$7,782,313", "-$2,097,636", "$10,408,207", "$2~
$ `Basketball Expenses`  <chr> "$16,176,262", "$14,127,220", "$10,733,046", "$~
$ `All Time Wins`        <dbl> 942, 875, 997, 941, 960, 960, 796, 926, 872, 91~
$ `Average Ticket Price` <dbl> 33, 10, 83, 15, 71, 30, 68, 60, 71, 78, 49, 76,~
$ `Years Played`         <dbl> 130, 131, 144, 128, 131, 134, 131, 142, 132, 13~
$ `State GDP M`          <dbl> 2355960, 755698, 620696, 240534, 277817, 822670~
$ `GDP Per Capita`       <dbl> 78456, 69253, 61859, 59894, 54753, 69978, 54753~
```

```
head(cfb_primary)
```

```
# A tibble: 6 x 22
   Rank School     State    City       `Win Rate` Conference `Football Revenue`
  <dbl> <chr>      <chr>    <chr>      <chr>      <chr>      <chr>
1     1 Texas      Texas    Austin     64%        Big 12     $162,450,371
2     2 Georgia    Georgia  Athens     89%        SEC        $155,951,449
3     3 Michigan   Michigan Ann Arbor  79%        Big Ten    $131,403,197
4     4 Oklahoma   Oklahoma Norman     79%        Big 12     $131,068,327
5     5 Alabama    Alabama  Tuscaloosa 90%        SEC        $130,868,208
6     6 Ohio State Ohio     Columbus   90%        Big Ten    $109,176,080
# i 15 more variables: `Football Profit` <chr>, `Football Expenses` <chr>,
#   `Undergrad Enrollment` <dbl>, `Football Attendence` <dbl>,
#   `Stadium Size` <dbl>, `% Fraternity` <dbl>, `% Sorority` <dbl>,
#   `Basketball Revenue` <chr>, `Basketball Profit` <chr>,
#   `Basketball Expenses` <chr>, `All Time Wins` <dbl>,
#   `Average Ticket Price` <dbl>, `Years Played` <dbl>, `State GDP M` <dbl>,
```

```
#   `GDP Per Capita` <dbl>
```

```
names(cfb_primary)
```

```
 [1] "Rank"                "School"              "State"
 [4] "City"                "Win Rate"            "Conference"
 [7] "Football Revenue"    "Football Profit"     "Football Expenses"
[10] "Undergrad Enrollment" "Football Attendence" "Stadium Size"
[13] "% Fraternity"        "% Sorority"          "Basketball Revenue"
[16] "Basketball Profit"   "Basketball Expenses" "All Time Wins"
[19] "Average Ticket Price" "Years Played"       "State GDP M"
[22] "GDP Per Capita"
```

```
nrow(cfb_primary)
```

```
[1] 107
```

The dataset has consistent structure but requires cleaning to address teams that changed conferences.

Secondary Data Set 1: 2023 College Football Team Stats

Where the Data Came From:
This dataset was sourced from a publicly available spreadsheet titled **cfb23_secondary.csv.**

Relevance to the Questions:
The 2023 dataset allows evaluation of offensive and defensive performance (yardage, scoring, turnovers, time of possession, and more) and how these things directly relate with win rates. This supports my ongoing investigation into what metrics contribute to a program's dominance.

Imported Data Inspection:

```
cfb23_secondary <- read_csv("cfb23_secondary.csv")
head(cfb23_secondary)
```

```
# A tibble: 6 x 152
   ...1 `Off Rank` Team            Games `Win-Loss` `Off Plays` `Off Yards`
  <dbl> <chr>      <chr>           <chr> <chr>      <chr>       <chr>
1     0 1          LSU (SEC)       13    3-Oct      841         7065
```

```
2     1 2          Oregon (Pac-12)   14    2-Dec      951         7440
3     2 3          Oklahoma (Big 12) 13    3-Oct      974         6591
4     3 4          Liberty (CUSA)    14    13-1       973         6988
5     4 5          Georgia (SEC)     14    13-1       958         6951
6     5 6          North Texas (AAC) 12    7-May      911         5950
# i 145 more variables: `Off Yards/Play` <chr>, `Off TDs` <chr>,
#   `Off Yards per Game` <chr>, `Def Rank` <chr>, `Def Plays` <chr>,
#   `Yards Allowed` <chr>, `Yards/Play Allowed` <chr>, `Off TDs Allowed` <chr>,
#   `Total TDs Allowed` <chr>, `Yards Per Game Allowed` <chr>,
#   `3rd Down Rank` <chr>, `3rd Attempts` <chr>, `3rd Conversions` <chr>,
#   `3rd Percent` <chr>, `3rd Down Def Rank` <chr>, `Opp 3rd Conversion` <chr>,
#   `Opp 3rd Attempt` <chr>, `Opponent 3rd Percent` <chr>, ...
```

```r
nrow(cfb23_secondary)
```

```
[1] 134
```

Secondary Data Set 2: 2013 College Football Team Stats

Where the Data Came From:
This dataset was sourced from a publicly available spreadsheet **cfb13_secondary.csv.**

Relevance to the Questions:
The 2013 dataset allows comparison over time, looking at how programs have maintained or declined in performance across a decade. This adds depth to my analysis of long-term program success.

Imported Data Inspection:

```r
cfb13_secondary <- read_csv("cfb13_secondary.csv")
head(cfb13_secondary)
```

```
# A tibble: 6 x 146
  Team     Games   Win  Loss Off.Rank Off.Plays Off.Yards Off.Yards.Play Off.TDs
  <chr>    <dbl> <dbl> <dbl>    <dbl>     <dbl>     <dbl>          <dbl>   <dbl>
1 Akron (~    12     5     7      106       837      4104            4.9      30
2 Alabama~    13    11     2       33       826      5903           7.15      58
3 Arizona~    13     8     5       31      1030      5960           5.79      52
4 Arizona~    14    10     4       32      1102      6402           5.81      64
5 Arkansa~    12     3     9       99       775      4286           5.53      29
6 Arkansa~    13     8     5       65       947      5301            5.6      45
```

```
# i 137 more variables: Total.TDs <dbl>, Off.Yards.per.Game <dbl>,
#   Def.Rank <dbl>, Def.Plays <dbl>, Yards.Allowed <dbl>,
#   Yards.Play.Allowed <dbl>, Off.TDs.Allowed <dbl>, Total.TDs.Allowed <dbl>,
#   Yards.Per.Game.Allowed <dbl>, First.Down.Rank <dbl>, First.Down.Runs <dbl>,
#   First.Down.Passes <dbl>, First.Down.Penalties <dbl>, First.Downs <dbl>,
#   First.Down.Def.Rank <dbl>, Opp.First.Down.Runs <dbl>,
#   Opp.First.Down.Passes <dbl>, Opp.First.Down.Penalties <dbl>, ...
```

```
nrow(cfb13_secondary)
```

```
[1] 111
```

Progress on Data Wrangling and Early EDA

In terms of the cleaning process I imported all three datasets successfully. I then began standardizing the team names across the years while also matching the conference names between primary and secondary datasets.

Early Visualizations:

```
library(rvest)
```

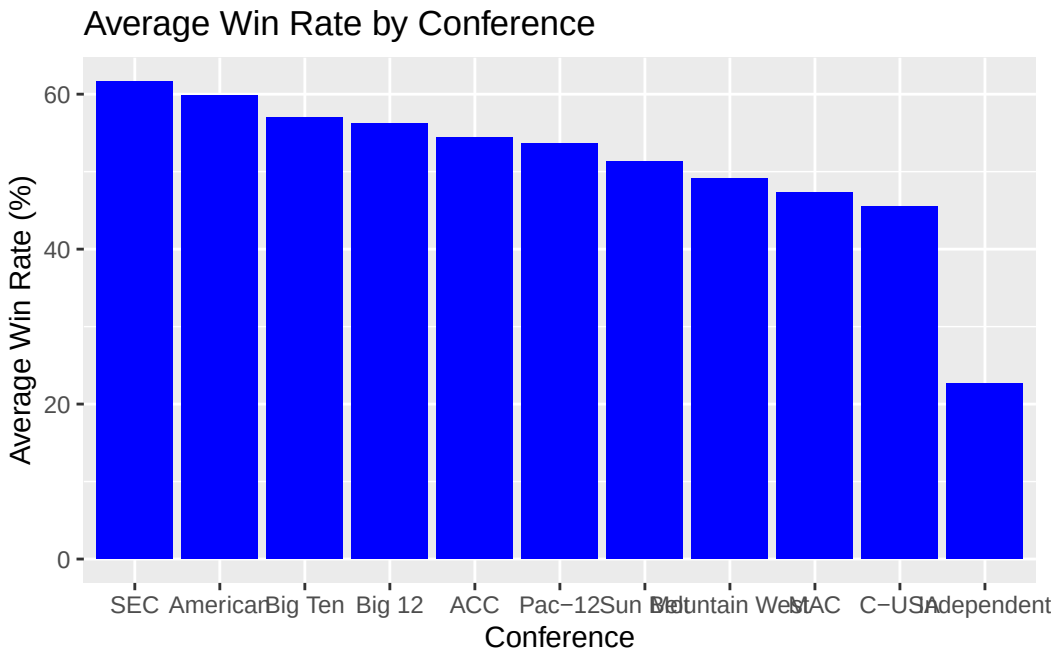
```
Attaching package: 'rvest'

The following object is masked from 'package:readr':

    guess_encoding
```

```
library(tidyverse)

url <- "https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_programs"
page <- read_html(url)

fbs_table <- page %>%
  html_table() %>%
  .[[1]]

fbs_clean <- fbs_table %>%
```

```
  select(
    School = `School`,
    Nickname = `Nickname`,
    City = `City`,
    State = `State [a]`,
    Enrollment = `Enrollment`,
    Conference = `CurrentConference[b]`
  )
```

In this next part, I made a plot for the Average Win Rate by Conference, as you can see in the plot, the larger and more historic conferences lead the way by percentages.

```
library(tidyverse)
cfb_primary <- read_csv("College Football.csv")
cfb_primary <- cfb_primary %>%
  rename(WinRate = `Win Rate`) %>%
  mutate(WinRate = as.numeric(str_remove(WinRate, "%")))

avg_win_rate <- cfb_primary %>%
  group_by(Conference) %>%
  summarize(AverageWinRate = mean(WinRate, na.rm = TRUE))

ggplot(avg_win_rate, aes(x = reorder(Conference, -AverageWinRate), y = AverageWinRate)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Average Win Rate by Conference",
       x = "Conference",
       y = "Average Win Rate (%)")
```



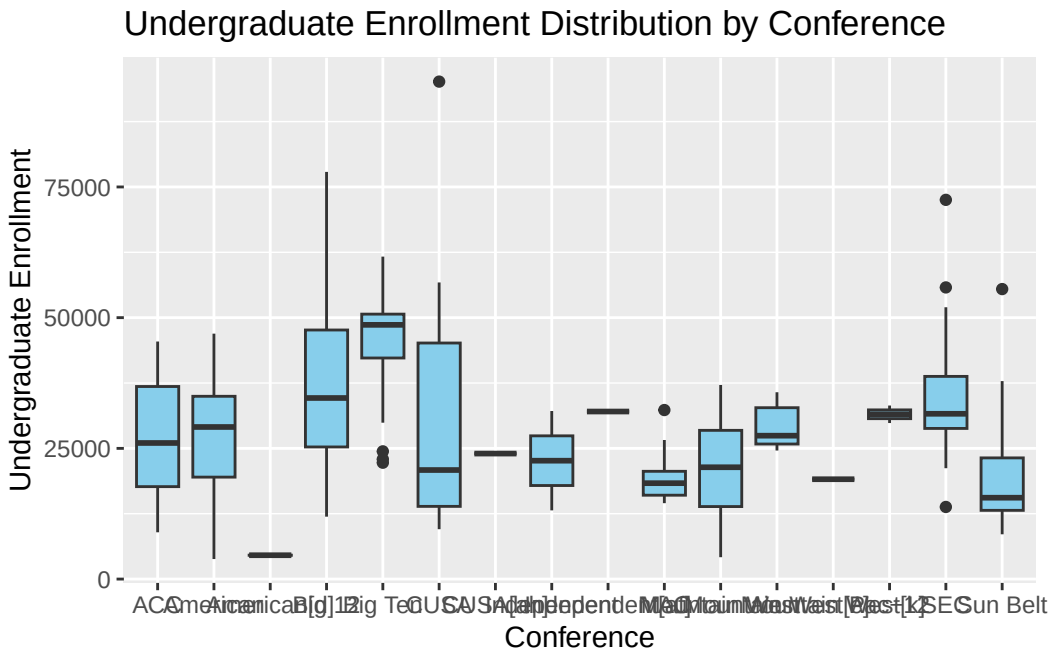
Web-Scraped Dataset

To explore how undergraduate enrollment varies across conferences in NCAA FBS programs, I made a boxplot using the web-scraped dataset from Wikipedia. This plot helps visualize the distribution and spread of enrollment sizes within each conference and it shows which conferences generally have larger or smaller schools.

```
library(rvest)
library(tidyverse)


url <- "https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_programs"
page <- read_html(url)

fbs_table <- page %>%
  html_table() %>%
  .[[1]]

# View columns
names(fbs_table)
```

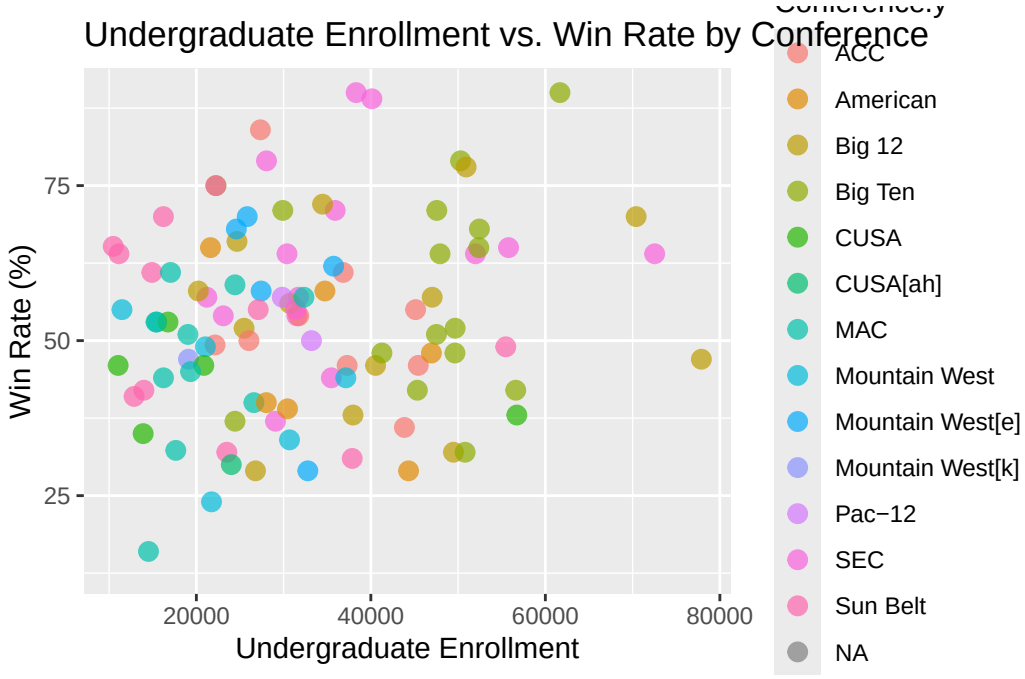
```
 [1] "School"               "Nickname"             "City"
 [4] "State [a]"            "Enrollment"           "CurrentConference[b]"
 [7] "FormerConferences"    "FirstYear"            "JoinedFBS"
[10] "First JoinedFBS"      "LeftFBS"
```

```r
fbs_clean <- fbs_table %>%
  select(
    School = `School`,
    Nickname = `Nickname`,
    City = `City`,
    State = `State [a]`,
    Enrollment = `Enrollment`,
    Conference = `CurrentConference[b]`
  )

fbs_clean <- fbs_clean %>%
  mutate(
    Enrollment = as.numeric(str_remove_all(Enrollment, ",")),
    Conference = as.factor(Conference)
  )

# Plot
ggplot(fbs_clean, aes(x = Conference, y = Enrollment)) +
  geom_boxplot(fill = "skyblue") +
  labs(
    title = "Undergraduate Enrollment Distribution by Conference",
    x = "Conference",
    y = "Undergraduate Enrollment"
  )
```



In this step, I joined my main football dataset with the data I scraped from Wikipedia so I could add each school's enrollment information. I then made a scatter plot to see if there is any pattern between a school's undergraduate enrollment and its football win rate, with the points colored by conference to compare across different groups.

```
cfb_augmented <- left_join(cfb_primary, fbs_clean, by = "School")

ggplot(cfb_augmented, aes(x = Enrollment, y = WinRate, color = Conference.y)) +
  geom_point(size = 3, alpha = 0.7) +
  labs(
    title = "Undergraduate Enrollment vs. Win Rate by Conference",
    x = "Undergraduate Enrollment",
    y = "Win Rate (%)",
    color = "Conference.y"
  )
```



Final Conclusion Paragraph

Working on this project gave me a chance to see how different factors shape the success of college football programs across conferences. By looking at win rates, enrollment sizes, and conference differences, I noticed that schools in conferences like the SEC and Big Ten often have higher win rates and bigger student bodies, which could give them an edge on the field. Using the all the data I collected I was able to tie together past performance with current trends in a way that made the numbers more meaningful in my opnion. This project also helped me get comfortable with cleaning data, joining datasets, creating visualizations, and even scraping data myself. Overall, it was interesting to see how the data lines up with what we often hear about college football, and I feel more confident analyzing real-world data after this project.

References

- NCAA Division I FBS Football Programs Wikipedia Page. (n.d.). Wikipedia. Retrieved from https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_programs
- College Football Statistics (2023 Season). Data pulled from https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019?resource=download

- College Football Statistics (2013 Season). Data pulled from https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019?resource=download
- Vergnolle, C. (n.d.). College Football Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/cvergnolle/football-5

Code Appendix

```
# Style Guide: Hadley Wickham's Tidyverse Style Guide

# Load Libraries
library(tidyverse)
library(stringr)
library(tidyr)


cfb_primary <- read_csv("College Football.csv", show_col_types = FALSE) %>%
  mutate(
    WinRate = as.numeric(str_remove(`Win Rate`, "%")),
    FootballRevenue = as.numeric(str_remove_all(`Football Revenue`, "[$,]")),
    FootballProfit = as.numeric(str_remove_all(`Football Profit`, "[$,]")),
    FootballExpenses = as.numeric(str_remove_all(`Football Expenses`, "[$,]"))
  )


cfb23_secondary <- read_csv("cfb23_secondary.csv", show_col_types = FALSE)
```

```
New names:
* `` -> `...1`
```

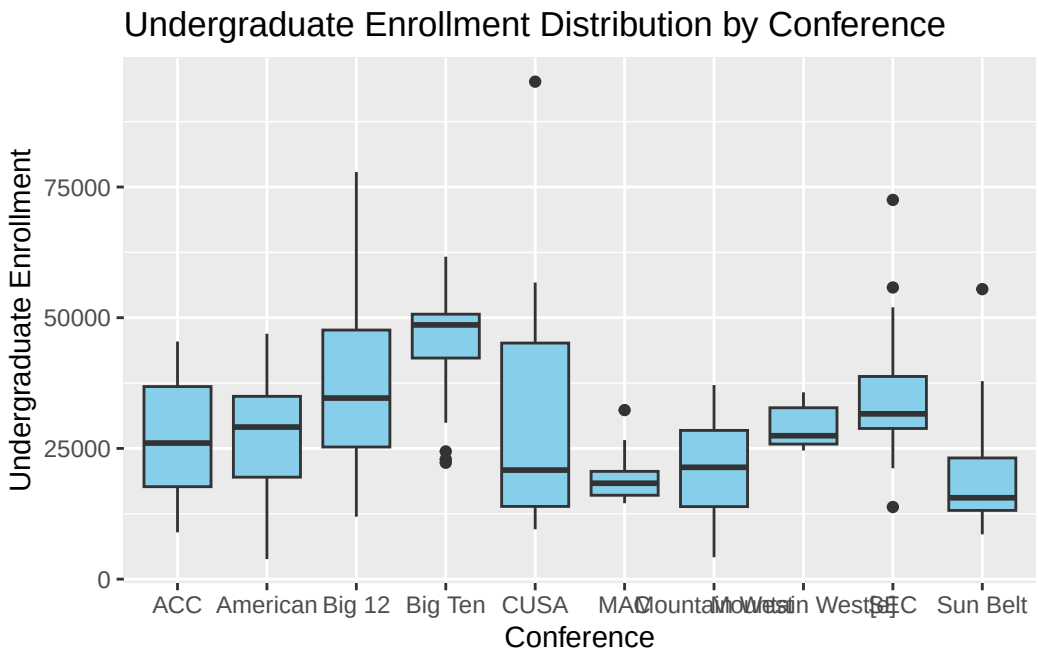
```
cfb23_secondary <- cfb23_secondary %>%
  mutate(TeamClean = str_remove(Team, " \\(.+\\)"))

cfb_combined <- cfb_primary %>%
  inner_join(cfb23_secondary, by = c("School" = "TeamClean"))


cfb_long <- cfb23_secondary %>%
  pivot_longer(
    cols = c(`Off Yards`, `Yards Allowed`),
```

```
    names_to = "Metric",
    values_to = "Yards"
  )


cfb_primary <- cfb_primary %>%
  mutate(
    Conference = str_replace_all(Conference, "Big 12", "Big Twelve")
  )

library(rvest)
library(tidyverse)


url <- "https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_programs"
page <- read_html(url)

fbs_table <- page %>%
  html_table() %>%
  .[[1]]

# View columns
names(fbs_table)
```

```
 [1] "School"               "Nickname"             "City"
 [4] "State [a]"            "Enrollment"           "CurrentConference[b]"
 [7] "FormerConferences"    "FirstYear"            "JoinedFBS"
[10] "First JoinedFBS"      "LeftFBS"
```

```
fbs_clean <- fbs_table %>%
  select(
    School = `School`,
    Nickname = `Nickname`,
    City = `City`,
    State = `State [a]`,
    Enrollment = `Enrollment`,
    Conference = `CurrentConference[b]`
  )

fbs_clean <- fbs_clean %>%
  mutate(
```

```
    Enrollment = as.numeric(str_remove_all(Enrollment, ",")),
    Conference = as.factor(Conference)
  )

# Filter to conferences with at least 5 schools
conf_counts <- fbs_clean %>%
  group_by(Conference) %>%
  tally() %>%
  filter(n >= 5)

fbs_filtered <- fbs_clean %>%
  filter(Conference %in% conf_counts$Conference)

# Plot
ggplot(fbs_filtered, aes(x = Conference, y = Enrollment)) +
  geom_boxplot(fill = "skyblue") +
  labs(
    title = "Undergraduate Enrollment Distribution by Conference",
    x = "Conference",
    y = "Undergraduate Enrollment"
  )
```



```r
library(rvest)
library(tidyverse)


url <- "https://en.wikipedia.org/wiki/List_of_NCAA_Division_I_FBS_football_programs"
page <- read_html(url)

fbs_table <- page %>%
  html_table() %>%
  .[[1]]

# View columns
names(fbs_table)
```

```
 [1] "School"               "Nickname"             "City"
 [4] "State [a]"            "Enrollment"           "CurrentConference[b]"
 [7] "FormerConferences"    "FirstYear"            "JoinedFBS"
[10] "First JoinedFBS"      "LeftFBS"
```

```r
fbs_clean <- fbs_table %>%
  select(
    School = `School`,
    Nickname = `Nickname`,
    City = `City`,
    State = `State [a]`,
    Enrollment = `Enrollment`,
    Conference = `CurrentConference[b]`
  )
```