Exploring_data01

Ian Vert

8/3/2021

With the start of the College Football Season soon approaching here is a beginning of data analysis of NCAA Football statistics.

This dataset is a combination of data from collegefootballdata.com

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.3     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.0     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(patchwork)
```

```
total_cfb <- read.csv('CFB2010_CFB2019.csv')
```

```
head(total_cfb[,c(1:9)])
```

```
##   Year          team    conference total.games total.wins total.losses
## 1 2010     Air Force Mountain West          13          9            4
## 2 2010         Akron  Mid-American          12          1           11
## 3 2010       Alabama           SEC          13         10            3
## 4 2010       Arizona        Pac-10          13          7            6
## 5 2010 Arizona State        Pac-10          12          6            6
## 6 2010      Arkansas           SEC          13         10            3
##   firstDowns fourthDownConversions fourthDowns
## 1        286                    18          30
## 2        175                     5          12
## 3        287                     9          14
## 4        308                     9          18
## 5        255                     8          14
## 6        291                     9          19
```

Here is a small look at what the dataframe looks like. This dataframe contains data from 2010 to 2019, 2020 data was not included as it was not a normal season.

The first Key Question that I would want to investigate is what is the most winningest conference, however number of conferences changed throughout the 2010s, this question had to be reframed with respect to starting with the College Football Playoff era starting in 2014.

Key Question: What conference is the is the most winningest in the College Football Playoff era?

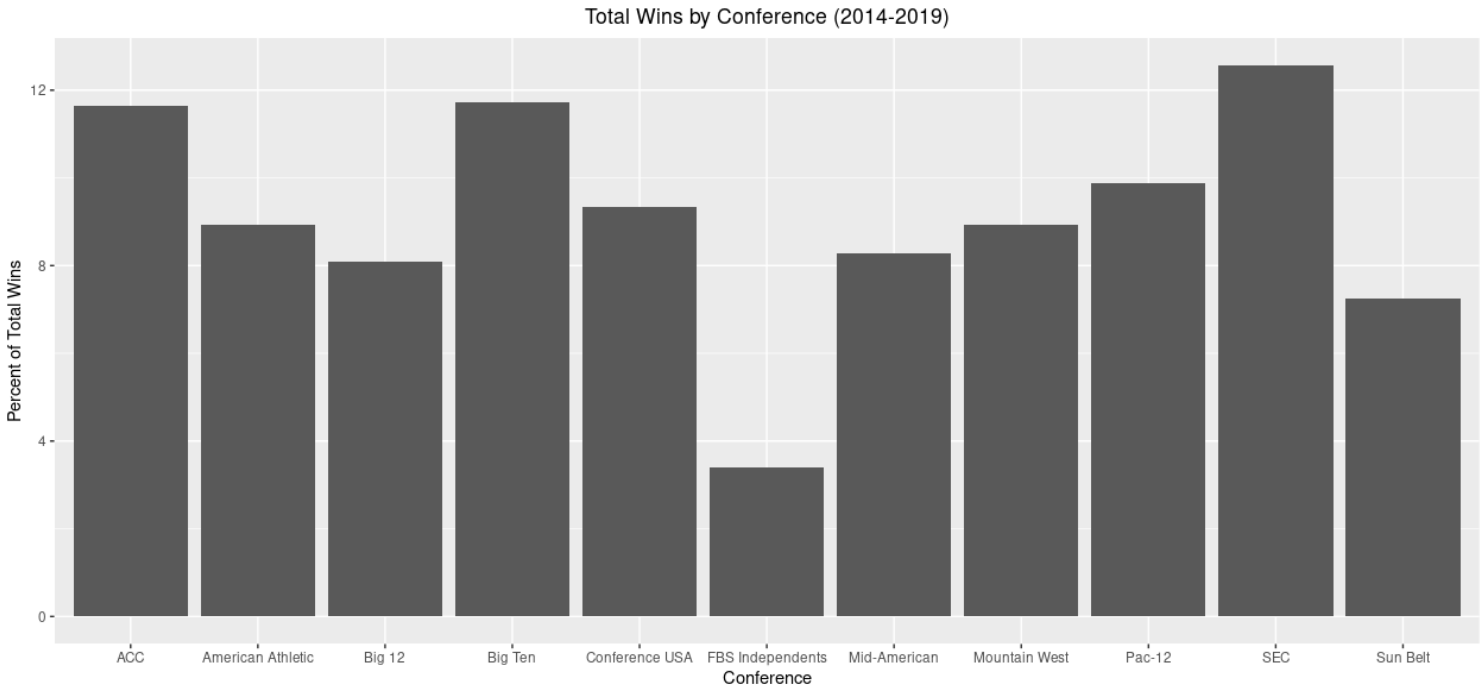
```
era_playoff <- total_cfb %>% filter(Year %in% (2014:2019))

head(era_playoff[,c(1:9)])
```

```
##   Year              team    conference total.games total.wins total.losses
## 1 2014         Air Force Mountain West          13         10            3
## 2 2014             Akron  Mid-American          12          5            7
## 3 2014           Alabama           SEC          14         12            2
## 4 2014 Appalachian State      Sun Belt          12          7            5
## 5 2014           Arizona        Pac-12          14         10            4
## 6 2014     Arizona State        Pac-12          13         10            3
##   firstDowns fourthDownConversions fourthDowns
## 1        287                   20          28
## 2        259                    7          21
## 3        340                   10          13
## 4        283                   14          23
## 5        346                   16          26
## 6        301                   10          23
```

```
conference_wins <- era_playoff  %>% group_by(conference) %>%
          mutate(Sum_wins = sum(total.wins)) %>%
          select(conference, Sum_wins) %>% unique()


conference_wins$Percent_wins <- conference_wins$Sum_wins/sum(conference_wins$Sum_wins) *
100

conference_wins
```

```
## # A tibble: 11 × 3
## # Groups:   conference [11]
##    conference        Sum_wins Percent_wins
##    <chr>                <int>        <dbl>
##  1 Mountain West          465         8.95
##  2 Mid-American           431         8.30
##  3 SEC                    655        12.6
##  4 Sun Belt               360         6.93
##  5 Pac-12                 515         9.92
##  6 FBS Independents       177         3.41
##  7 Big 12                 422         8.13
##  8 ACC                    606        11.7
##  9 American Athletic      465         8.95
## 10 Conference USA         486         9.36
## 11 Big Ten                611        11.8
```

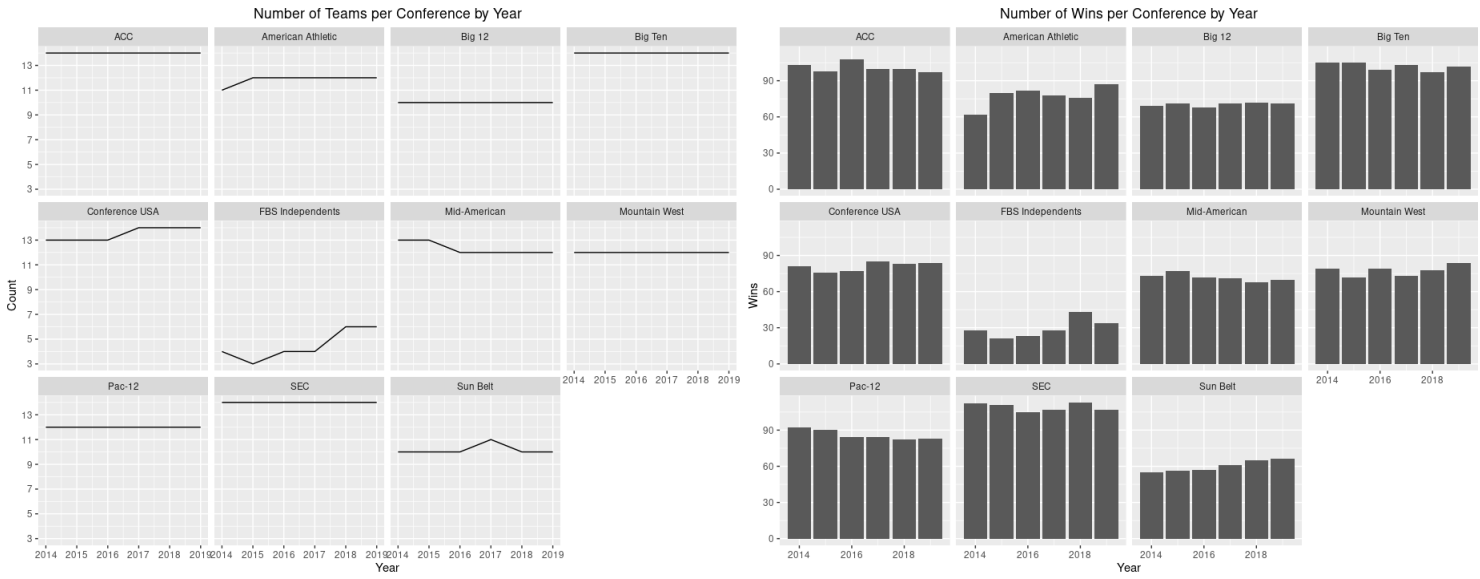
```
p <- conference_wins %>% ggplot(mapping= aes(x = conference, y = Percent_wins)) +
geom_bar(stat = 'identity') +
labs(x = 'Conference', y = 'Percent of Total Wins', title = 'Total Wins by Conference
(2014-2019)') + theme(plot.title = element_text(hjust = 0.5))

plot(p)
```



This bar graph shows the Percent of total wins per conference from the start of the College Football Playoff in 2014. This graph shows that the SEC has the most wins. However, this graph is misleading as each conference does not have an equal number of teams.

```
teams_per_conf <- era_playoff %>% count(Year,conference) %>% ggplot(mapping = aes(x =
Year , y = n)) +
   geom_line() +
  facet_wrap(~conference) +
  scale_y_continuous(breaks = c(3,5,7,9,11,13)) +
  labs(x = 'Year',
       y = 'Count',
       title = 'Number of Teams per Conference by Year') + theme(plot.title =
element_text(hjust = 0.5))


wins_per_conf <- era_playoff %>% select(Year, conference, total.wins) %>% ggplot(mapping
= aes(x = Year , y = total.wins)) +
   geom_bar(stat = 'identity') +
  facet_wrap(~conference) +
  labs(x = 'Year',
       y = 'Wins',
       title = 'Number of Wins per Conference by Year') + theme(plot.title =
element_text(hjust = 0.5))
```

```
patch1 <- teams_per_conf + wins_per_conf

patch1
```



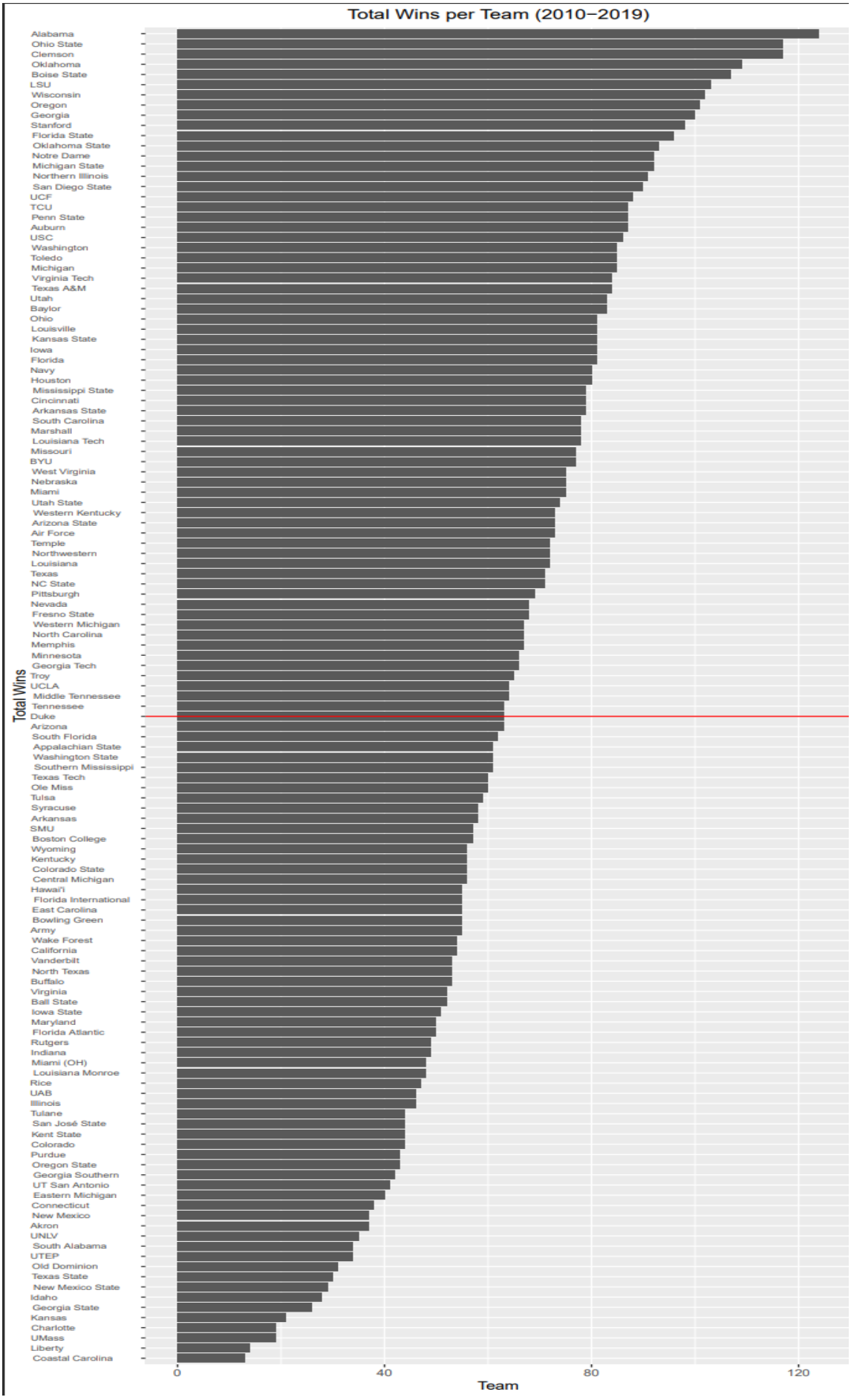
The graph above puts better context on the investigation of the winningest conference by viewing number of teams per conference by year alongside wins of each conference by year. Certain conferences such as the Big-12, Big-10, ACC, Pac-12, and the SEC usually known as the "Power-Five" had consistent number of team in their conference but showed slight variability in the number of wins per season. While conferences like Conference-USA and the Independents showed an increase in wins as the teams in the conference increased.

Key Question: What are the winningest Teams from 2010-2019?

After looking at the winningest conferences in the College Football Playoff era, looking at specific teams is the next step.

```
df_team_wins <- total_cfb %>% select(team | conference | total.wins)

total_team_wins <- df_team_wins %>% group_by(team) %>% mutate(Sum_team_wins =
sum(total.wins)) %>% select(team, Sum_team_wins) %>% arrange(Sum_team_wins) %>% unique()

tail(total_team_wins)

## # A tibble: 6 × 2
## # Groups:   team [6]
##   team        Sum_team_wins
##   <chr>               <int>
## 1 LSU                   103
## 2 Boise State           107
## 3 Oklahoma              109
```

```
## 4 Clemson                117
## 5 Ohio State             117
## 6 Alabama                124

total_team_wins$team <- factor(total_team_wins$team, levels = total_team_wins$team)


p <- total_team_wins %>% ggplot(mapping = aes(x = team, y = Sum_team_wins))+
  geom_bar(stat = 'Identity') +
  labs(x = 'Total Wins',
      y = 'Team',
     title = 'Total Wins per Team (2010-2019)') +

  theme(axis.text.y = element_text(size = 7, hjust = -0.05)) + coord_flip()  +
   geom_vline(aes(xintercept = median(Sum_team_wins)), color = 'red') +
  theme(plot.title = element_text(hjust = 0.5))


plot(p)
```


Total Wins per Team (2010–2019)
Alabama
Ohio State
Clemson
Oklahoma
Boise State
LSU
Wisconsin
Oregon
Georgia
Stanford
Florida State
Oklahoma State
Notre Dame
Michigan State
Northern Illinois
San Diego State
UCF
TCU
Penn State
Auburn
USC
Washington
Toledo
Michigan
Virginia Tech
Texas A&M
Utah
Baylor
Ohio
Louisville
Kansas State
Iowa
Florida
Navy
Houston
Mississippi State
Cincinnati
Arkansas State
South Carolina
Marshall
Louisiana Tech
Missouri
BYU
West Virginia
Nebraska
Miami
Utah State
Western Kentucky
Arizona State
Air Force
Temple
Northwestern
Louisiana
Texas
NC State
Pittsburgh
Nevada
Fresno State
Western Michigan
North Carolina
Memphis
Minnesota
Georgia Tech
Troy
UCLA
Middle Tennessee
Tennessee
Duke
Arizona
South Florida
Appalachian State
Washington State
Southern Mississippi
Texas Tech
Ole Miss
Tulsa
Syracuse
Arkansas
SMU
Boston College
Wyoming
Kentucky
Colorado State
Central Michigan
Hawai'i
Florida International
East Carolina
Bowling Green
Army
Wake Forest
California
Vanderbilt
North Texas
Buffalo
Virginia
Ball State
Iowa State
Maryland
Florida Atlantic
Rutgers
Indiana
Miami (OH)
Louisiana Monroe
Rice
UAB
Illinois
Tulane
San José State
Kent State
Colorado
Purdue
Oregon State
Georgia Southern
UT San Antonio
Eastern Michigan
Connecticut
New Mexico
Akron
UNLV
South Alabama
UTEP
Old Dominion
Texas State
New Mexico State
Idaho
Georgia State
Kansas
Charlotte
UMass
Liberty
Coastal Carolina
Total Wins
0
40
80
120
Team

The graph above This graph shows how many games each team won, from the past decade (2010-2019).

The red line shows the median number of wins.

This shows the dominance of teams that are regulars in the College Football Playoffs such as Alabama, Ohio State, and Clemson. The graph also shows teams that commonly do not have a chance at to make the College Football Playoffs, such as Boise State, Toledo, and UCF as they dominate conferences that part of what is commonly refered to as "Group of 5".

This shows that as there is conference realignment there is an opportunity to restructure the conferences to make them more equal in size and talent to give more teams an opportunity at the playoff.

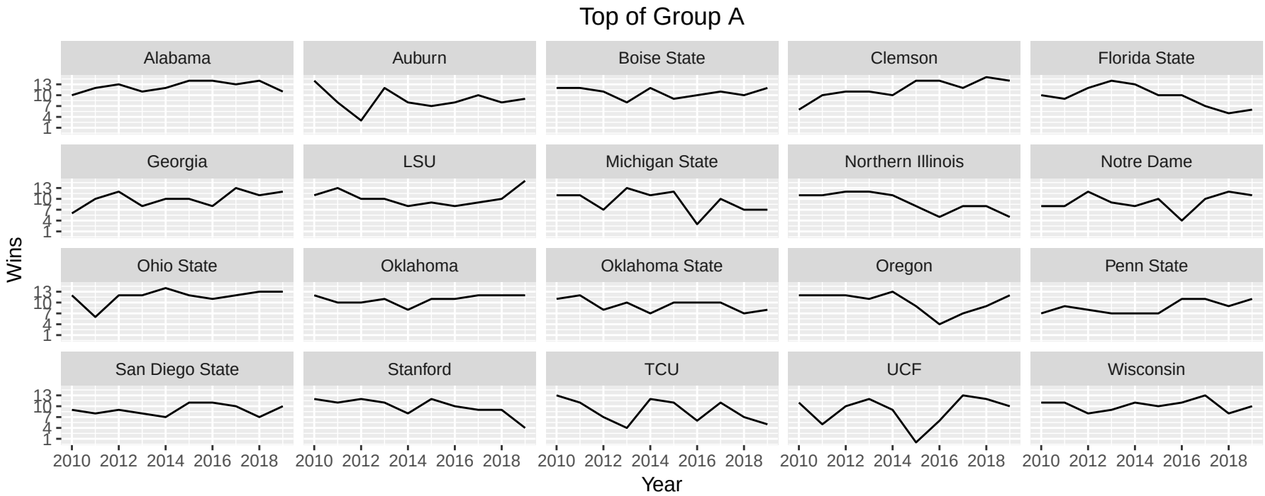
```
rank_teams <- total_team_wins %>% mutate(Decade_Rank =
                              case_when(Sum_team_wins >= 64 ~ 'Above_team_median',
                                        Sum_team_wins  <  64 ~ 'Below_team_median'))


total_cfb <- merge(x = total_cfb , y = rank_teams %>% select(team,Sum_team_wins,
Decade_Rank), by ='team')

total_cfb <-total_cfb%>%
select(Year,team,conference:tacklesForLoss,Sum_team_wins,Decade_Rank,conferenceGames.game
s:location.longitude) %>% arrange(Year)

group_A <- total_cfb %>% filter(Decade_Rank == 'Above_team_median') %>%
arrange(desc(Sum_team_wins),Year)

group_B <- total_cfb %>% filter(Decade_Rank == 'Below_team_median') %>%
arrange(desc(Sum_team_wins),Year)
```

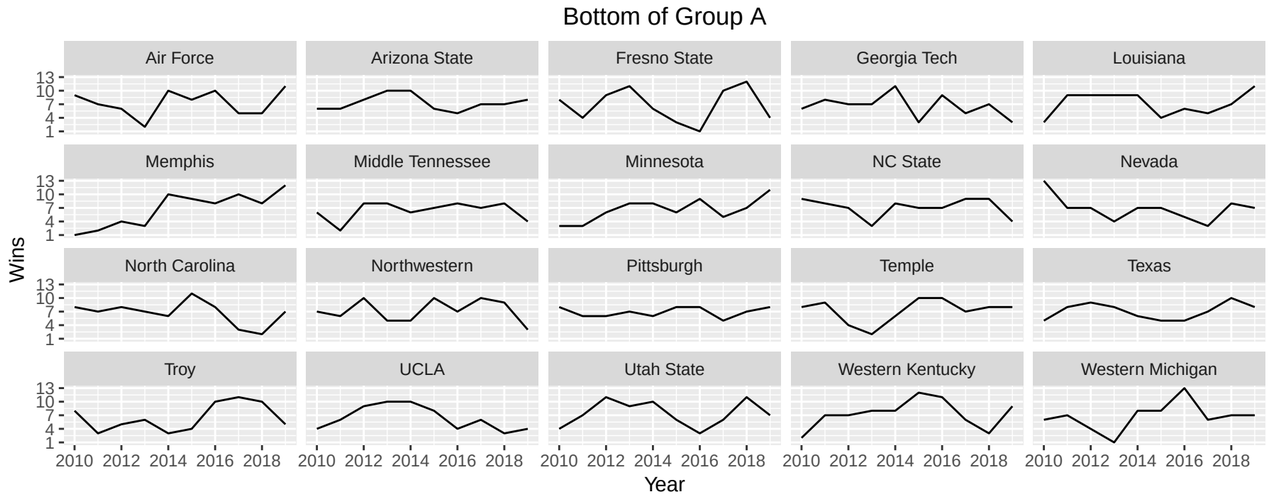
Two new dataframes have been created. The teams will be split into two groups Group A , and Group B. Group A are the teams above the median line in the previous figure, and Group B are the teams below the median line the in the previous figure.

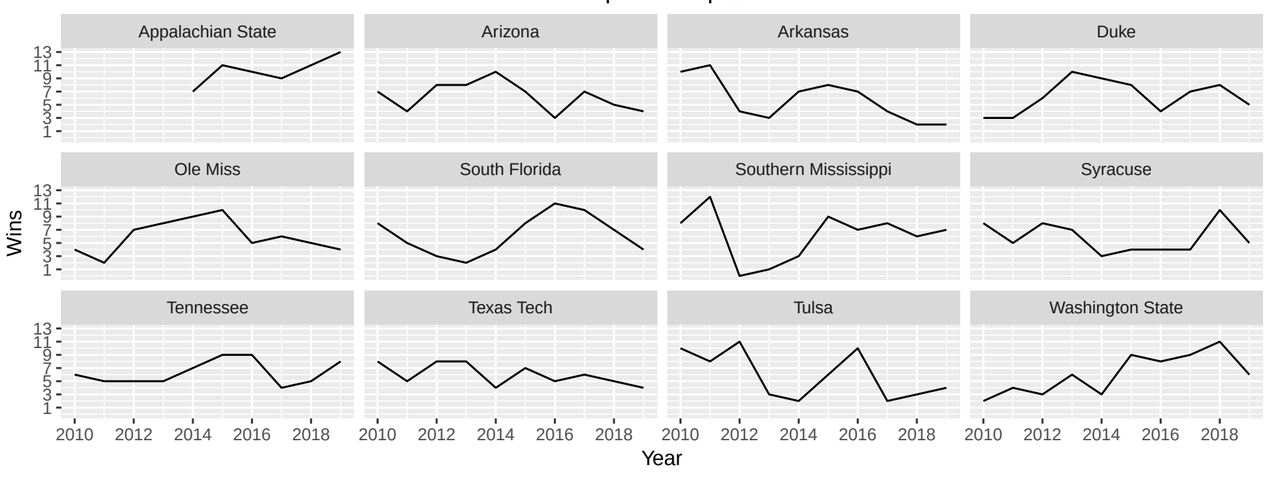
Key Question: What is the difference in the variability of wins per year between the top of Group A to the bottom of Group A? What is the difference in the variability of wins per year between Group A and Group B?

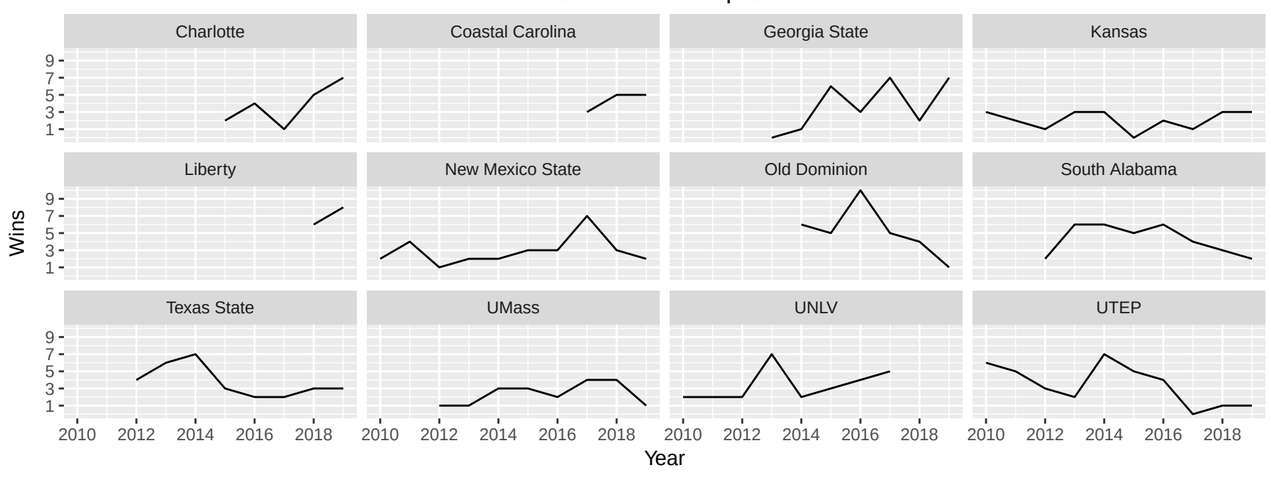
```
plot_group_A_top <- group_A[c(1:200),] %>% ggplot(mapping = aes(x= Year, y = total.wins))
+ geom_line() +
  scale_x_continuous(breaks = seq(2010,2019,2)) +
  scale_y_continuous(breaks=seq(1,17,3)) + facet_wrap(~team) +
  labs(x = 'Year', y = 'Wins', title = 'Top of Group A',) +
  theme(plot.title = element_text(hjust = 0.5))

plot_group_A_bottom <- (group_A %>% arrange((Sum_team_wins),Year))[c(1:200),] %>%
ggplot(mapping = aes(x= Year, y = total.wins)) + geom_line() +
   scale_x_continuous(breaks = seq(2010,2019,2)) +
```

```
  scale_y_continuous(breaks=seq(1,17,3)) + facet_wrap(~team) +
  labs(x = 'Year', y = 'Wins', title = 'Bottom of Group A',) +
  theme(plot.title = element_text(hjust = 0.5))


plot_group_B_top <- group_B[c(1:116),] %>% ggplot(mapping = aes(x= Year, y = total.wins))
+ geom_line(aes(group=team)) +
  scale_x_continuous(breaks = seq(2010,2019,2)) +
  scale_y_continuous(breaks=seq(1,17,2)) + facet_wrap(~team) +
  labs(x = 'Year', y = 'Wins', title = 'Top of Group B',) +
  theme(plot.title = element_text(hjust = 0.5))

plot_group_B_bottom <-(group_B %>% arrange((Sum_team_wins),Year))[c(1:85),] %>%
ggplot(mapping = aes(x= Year, y = total.wins)) + geom_line() +
  scale_x_continuous(breaks = seq(2010,2019,2)) +
  scale_y_continuous(breaks=seq(1,17,2)) + facet_wrap(~team) +
  labs(x = 'Year', y = 'Wins', title = 'Bottom of Group B ',) +
  theme(plot.title = element_text(hjust = 0.5))

group_patch <-plot_group_A_top + plot_group_A_bottom + plot_group_B_top +
plot_group_B_bottom + plot_layout(widths=1)

group_patch
```


Top of Group A
Alabama
Auburn
Boise State
Clemson
Florida State
Georgia
LSU
Michigan State
Northern Illinois
Notre Dame
Ohio State
Oklahoma
Oklahoma State
Oregon
Penn State
San Diego State
Stanford
TCU
UCF
Wisconsin
Wins
Year


Bottom of Group A
Air Force
Arizona State
Fresno State
Georgia Tech
Louisiana
Memphis
Middle Tennessee
Minnesota
NC State
Nevada
North Carolina
Northwestern
Pittsburgh
Temple
Texas
Troy
UCLA
Utah State
Western Kentucky
Western Michigan
Wins
Year


Top of Group B
Appalachian State
Arizona
Arkansas
Duke
Ole Miss
South Florida
Southern Mississippi
Syracuse
Tennessee
Texas Tech
Tulsa
Washington State
Wins
Year


Bottom of Group B
Charlotte
Coastal Carolina
Georgia State
Kansas
Liberty
New Mexico State
Old Dominion
South Alabama
Texas State
UMass
UNLV
UTEP
Wins
Year

The Graph above show the number of wins per year for the Top and Bottom of each Group. The two graphs show an interesting picture of College Football. The variability in wins per year in Group B is more dramatic than Group A. An example would be Arizona which in 2011 had 4 wins and in 2016 had 3 wins but in 2014 hit their decade win high of 10 wins. Another example, South Florida had an increase of wins from 2013 to 2016 with 11 wins. But then preceded to decrease in wins since ending 2019 with 4 wins. While the team at the top of Group A: Alabama, had been very consistent with wins. The variability in wins increases from the top of Group A to the bottom of Group B. This shows the consistency of the best programs in college football such as Alabama, Ohio State, and Clemson.

Patterns to look at:

"V" shaped Recovery: Ohio State, Oklahoma, Michigan State, Notre Dame, and UCF.

Ohio State and Oklahoma show similar trajectories in that they had a significant drop in wins for a year, 2011 for Ohio State and 2014 for Oklahoma, but since those years both programs have had consistent wins per season.

The quick 'V' shaped recovery is shown in Michigan State, Notre Dame, and UCF. The most dramatic 'v' shaped recovery was UCF which had 9 wins in 2014, 0 wins in 2015, 13 in 2017. The will be further exploring of the data to see if the stats show possible reasons for the 'V' shaped recovery.

Plateau

A plateau of more than two years is shown in Oklahoma State and Penn State. A reason for this plateau is not known from the data present but will be investigated in the future.

Yo-yo

Looking at both TCU and Northwestern throughout the decade there is no consistency just up-down-up-down pattern.

Quick Peak and Steep Fall

A Quick peak and fall can be seen with North Carolina, and Georgia State, which both within the decade had a jump of 4 wins or more in a year and then never reached that level of wins in the following years.

Note: The graph showing the Bottom of Group B shows teams with missing data such as Idaho, which left the FBS to go to the FCS and teams such as Coastal Carolina which moved from FCS to FBS.

Looking to the future of College Football:

Key Question: How would all 4 graphs above look for 2020-2029?

With the advent of NIL deals, imminent conference restructuring, and possible College Football expansion the future of College Football is exciting but uncertain.

If 'super' conferences with over 20 teams, visualizing the winningest conference would cease to be useful. With NIL deals, and teams gaining more exposure through a playoff expansion this may help teams to maintain consistency in recruiting and allow for more rapid increase in wins, like the 'v' shaped recovery exhibited in the last graph. However, the increased accessibility of the transfer portal may cause teams to have Quick peaks and fall as their best player leave for more consistent winning programs. With the changes to College Football a graph such as the third image for 2020-2029 could show teams more equal in wins or will exacerbate the inequality and increase the gap between the teams at the top of 'Group A' and the teams at the bottom of 'Group A'.

The patterns exhibited in the fourth image are due to the nature of college football, with consistent roster turnover and would not be fixed with conference restructuring. I would like to compare these patterns with a 10-year period of the NFL.

Part of the future exploration of this data set will be looking at what the future the College Football may look like.

Here I will save the new data frame for Exploring_data02.

```
write.csv(total_cfb,'CFB_Total.csv', row.names =  FALSE)
```