DataAppendix

Import packages

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(readxl)
library(skimr)
library(dplyr)
```

Read in table

```
football <- readxl::read_excel("football.xlsx")
skim(football)
```

Table 1: Data summary

Name	football
Number of rows	131

Number of columns	52
Column type frequency:	
character	1
numeric	51
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Team	0	1	9	33	0	131	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Games	0	1	12.76	0.79	10.00	12.00	13.00	13.00	15.00	
Wins	0	1	6.78	2.89	1.00	5.00	7.00	9.00	15.00	
Losses	0	1	5.98	2.30	0.00	4.00	6.00	7.00	11.00	
winPer	0	1	0.52	0.20	0.08	0.42	0.54	0.69	1.00	
Total_Points	0	1	362.27	103.30	150.00	285.00	360.00	427.50	616.00	
Points_Per_Game	0	1	28.14	6.96	12.50	23.35	28.20	32.75	46.10	
Points_Allowed	0	1	334.45	70.02	166.00	292.00	331.00	378.00	534.00	
Avg_Points_per_Game_Allowed	0	1	26.33	5.86	12.80	22.35	26.70	29.75	44.50	
Off_Yards	0	1	5025.44	973.11	2737.00	4346.00	4928.00	5668.00	7517.00	
Off_Yards/Play	0	1	5.70	0.73	3.94	5.23	5.70	6.17	7.28	
Off_TDs	0	1	43.07	13.99	14.00	32.00	43.00	53.00	78.00	
Off_Yards_per_Game	0	1	391.89	60.79	228.10	351.30	389.40	435.90	525.50	
Yards_Allowed	0	1	4836.30	662.83	3184.00	4324.00	4816.00	5265.00	6433.00	
Yards/Play_Allowed	0	1	5.55	0.59	3.99	5.14	5.54	5.88	7.42	
Off_TDs_Allowed	0	1	39.40	9.62	15.00	34.00	40.00	46.00	70.00	
Total_TDs_Allowed	0	1	41.53	9.96	16.00	34.50	41.00	48.00	72.00	
Yards_Per_Game_Allowed	0	1	379.59	51.58	254.40	340.35	377.00	413.20	516.60	
Penalties	0	1	76.53	14.49	46.00	67.00	76.00	86.00	111.00	
Penalty_Yards	0	1	671.96	132.55	376.00	584.50	671.00	764.00	1007.00	
Penalty_Yards_Per_Game	0	1	52.72	10.23	29.69	45.28	53.64	60.04	77.46	
Pass_Yards/Attempt	0	1	7.37	1.03	4.94	6.72	7.41	8.12	10.53	
Yards/Completion	0	1	12.25	1.86	9.19	11.16	12.11	13.00	22.34	
Pass_Yards_Per_Game	0	1	233.01	52.58	70.50	206.30	232.40	266.65	369.80	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Yards/Attempt_Allowed	0	1	7.17	0.80	5.27	6.56	7.16	7.67	9.74	
Yards/Completion_Allowed	0	1	11.90	1.06	8.99	11.11	11.91	12.55	14.71	
Pass_Yards_Per_Game_Allowed	0	1	227.16	32.56	156.20	206.50	224.80	250.40	294.70	
Rushing_Def_Rank	0	1	65.98	37.96	1.00	33.50	66.00	98.50	131.00	
Opp_Rush_Attempts	0	1	465.21	47.01	354.00	436.00	465.00	498.00	571.00	
Opp_Rush_Yards_Allowed	0	1	1936.81	430.87	874.00	1632.00	1949.00	2226.50	3002.00	
Yds/Rush_Allowed	0	1	4.14	0.70	2.47	3.70	4.05	4.54	6.12	
Opp_Rush_Touchdowns_Allowed	0	1	19.42	6.71	4.00	15.00	19.00	23.00	39.00	
Rush_Yards_Per_Game_Allowed	0	1	152.40	35.70	77.10	126.60	149.60	174.45	245.10	
Rushing_Off_Rank	0	1	65.99	37.96	1.00	33.50	66.00	98.50	131.00	
Rush_Attempts	0	1	470.85	80.45	295.00	416.00	470.00	519.00	802.00	
Rush_Yds	0	1	2039.58	626.91	758.00	1541.50	2003.00	2464.00	4247.00	
Yards/Rush	0	1	4.26	0.79	2.09	3.63	4.23	4.92	6.00	
Rushing_Yards_per_Game	0	1	158.89	45.19	63.20	124.95	155.70	190.70	326.70	
Touchdowns_Allowed	0	1	41.51	9.97	16.00	34.50	41.00	48.00	72.00	
Opp_Field_Goals_Made	0	1	14.80	3.27	7.00	13.00	15.00	17.00	23.00	
Touchdowns	0	1	45.45	14.15	16.00	34.00	45.00	55.00	79.00	
Field_Goals	0	1	14.87	4.79	3.00	11.50	14.00	18.00	29.00	
Sacks	0	1	27.16	10.82	5.00	19.00	27.00	35.00	59.00	
Sack_Yards	0	1	178.61	72.99	24.00	126.00	174.00	224.00	399.00	
Average_Sacks_per_Game	0	1	2.14	0.88	0.38	1.48	2.08	2.84	4.92	
Tackle_for_Loss_Yards	0	1	295.44	70.35	166.00	245.00	287.00	343.00	510.00	
Total_Tackle_For_Loss	0	1	74.56	15.00	47.00	62.50	74.00	86.00	121.00	
Tackle_For_Loss_Per_Game	0	1	5.84	1.12	3.60	5.00	5.80	6.65	9.30	
Turnovers_Gain	0	1	18.77	4.84	8.00	15.50	19.00	22.00	32.00	
Turnovers_Lost	0	1	18.20	4.79	7.00	15.00	18.00	21.00	32.00	
Turnover_Margin	0	1	0.57	7.00	-19.00	-4.00	1.00	5.50	22.00	
Avg_Turnover_Margin_per_Game	0	1	0.03	0.55	-1.58	-0.33	0.07	0.41	1.57	

Variables

No missing values in any column.

Team is type chr string of the college/team name. Every FBS school is included here.

Conference might be a variable that we want, but it likely doesn't have too much of an effect. Type chr string for each school.

Wins contains the number of wins a team had last season. Has type integer with range 0-16, which makes sense.

winPer is the win percentage of a team last season. This might be more helpful than wins because some stats may be skewed one way or another if a team played more or less games than another. Its a numeric value with range 0-1, which makes sense. No missing values.

Points_per_game is the amount of points per game a team scores. This also is better than total points for the same reason. This is a numeric value with range 10-50, which makes sense too.

Avg_points_per_game_allowed is the same as points per game, but for the opposing teams a team has played. This is a numeric value with range 10-45, so this makes sense.

Off_yards/play is also numeric, and has range 3.5-7.5, which is valid. This is the amount of offensive yards per play a team has.

Yards/play_allowed is the same as Off_yards/play but for defense. Numeric with range 3.5-7.5.

Penalty_Yards_per_game could be important because more disciplined teams might be more capable of winning games. This is also a numeric with range 25-80, which makes sense.

AvgTurnoverMarg could be very useful because teams who win, usually win the turnover battle. This is numeric with range -2-2, negative values being a negative margin, and vice versa.

All of the variable types and ranges make a lot of sense to us, and there are no issues it seems.

Pressing issues

1. Separating team name and conference name, and removing spaces
2. Replacing all spaces in column names with underscores
3. Selecting columns we think we need and removing all others

Data Cleaning

```
football <- football |>
  mutate(AvgTurnoverMarg = Avg_Turnover_Margin_per_Game) |>
  select(-Avg_Turnover_Margin_per_Game) |>
  separate_wider_delim(Team, names = c("TeamName", "conference"), delim = "(", too_many =
  mutate(TeamName = str_replace_all(TeamName, "_", "")) |>
  mutate(conference = str_replace_all(conference, "/", "_")) |>
```

```
  rename_all(~str_replace_all(., "/", "_per_")) |>
  mutate(conference = str_remove_all(conference, "\\)")) |>
  select(TeamName, conference, Wins, winPer, Points_Per_Game, Avg_Points_per_Game_Allowed,
```

Skim

```
skim(football)
```

Table 4: Data summary

Name	football
Number of rows	131
Number of columns	10
_____	
Column type frequency:	
character	2
numeric	8
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
TeamName	0	1	3	18	0	130	0
conference	0	1	3	15	0	13	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Wins	0	1	6.78	2.89	1.00	5.00	7.00	9.00	15.00	
winPer	0	1	0.52	0.20	0.08	0.42	0.54	0.69	1.00	
Points_Per_Game	0	1	28.14	6.96	12.50	23.35	28.20	32.75	46.10	
Avg_Points_per_Game_Allowed	0	1	26.33	5.86	12.80	22.35	26.70	29.75	44.50	
Penalty_Yards_Per_Game	0	1	52.72	10.23	29.69	45.28	53.64	60.04	77.46	
AvgTurnoverMarg	0	1	0.03	0.55	-1.58	-0.33	0.07	0.41	1.57	
Off_Yards_per_Play	0	1	5.70	0.73	3.94	5.23	5.70	6.17	7.28	
Yards_per_Play_Allowed	0	1	5.55	0.59	3.99	5.14	5.54	5.88	7.42	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist

There are no missing values, and the number of variables is now 8, which is much more manageable. These are also the exact variables that should be relevant to predicting wins or win percentage.

Everything seems to have worked, and they are in our desirable data types.