Vikings_EDA

Exploring Vikings NFL Data

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
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 becom
```

```r
vikes_data <-read_csv("final_vikings_data.csv")
```

```
Rows: 27612 Columns: 255
-- Column specification --------------------------------------------------------
Delimiter: ","
chr   (74): home_team, away_team, posteam, posteam_type, defteam, side_of_fi...
dbl  (147): play_id, game_id, yardline_100, quarter_seconds_remaining, half_...
lgl   (32): lateral_receiver_player_id, lateral_receiver_player_name, latera...
date   (1): game_date
time   (1): time

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
## Getting to know the data
dim(vikes_data)
```

```
[1] 27612   255
```

```r
vikes_data<-vikes_data|>
  filter(!is.na(posteam) & posteam=="MIN") # I'm going to get only the plays where the vik
```

Contingency Tables and Numeric Summaries

First, visualize vikings play type per down.

```r
table(vikes_data$down, vikes_data$play_type)
```

```

    extra_point field_goal kickoff no_play pass punt qb_kneel qb_spike  run
  1           0          9       0     250 1991    0       67        9 2340
  2           0          6       0     193 1870    0       34        1 1415
  3           0         10       0     189 1642    0       25        1  419
  4           0        289       0      57   94  726        1        0   47
```

Seems like the Vikings are more likely to pass over run on later downs.

Let's look at counts and ratios of pass and run plays per year.

```r
library(lubridate)
vikes_table_1 <-vikes_data|>
  mutate(year = year(game_date)) |>
  group_by(year) |>
  summarize(run_count=sum(play_type=="run",na.rm=TRUE ), pass_count = sum(play_type=="pass
  mutate(run_ratio = run_count/(run_count+pass_count),pass_ratio = pass_count/(run_count+p

vikes_table_1
```

```
# A tibble: 10 x 5
    year run_count pass_count run_ratio pass_ratio
   <dbl>     <int>      <int>     <dbl>      <dbl>
 1  2009       419        547     0.434      0.566
 2  2010        439        540     0.448      0.552
 3  2011        440        548     0.445      0.555
 4  2012        499        560     0.471      0.529
 5  2013        415        590     0.413      0.587
 6  2014        402        563     0.417      0.583
 7  2015        435        476     0.477      0.523
 8  2016        369        616     0.375      0.625
 9  2017        505        589     0.462      0.538
10  2018        305        581     0.344      0.656
```

I want to create a table that shows average yards per play by year.

```
vikes_table_2 <- vikes_data |>
  mutate(year = year(game_date)) |>
  group_by(year) |>
  summarize(
    avg_yards = mean(yards_gained, na.rm = TRUE),
    yards_sd  = sd(yards_gained, na.rm = TRUE)
  )
```

```
vikes_table_2
```

```
# A tibble: 10 x 3
    year avg_yards yards_sd
   <dbl>     <dbl>    <dbl>
 1  2009      4.28     8.39
 2  2010      3.98     7.54
 3  2011      3.90     8.14
 4  2012      3.99     8.20
 5  2013      4.04     8.36
 6  2014      3.79     7.70
 7  2015      3.92     8.29
 8  2016      3.69     7.16
 9  2017      4.07     7.83
10  2018      4.22     7.83
```

Cool! Vikings were averaging a high 4.3 yards per play in 2009. The Vikings were 12-4.

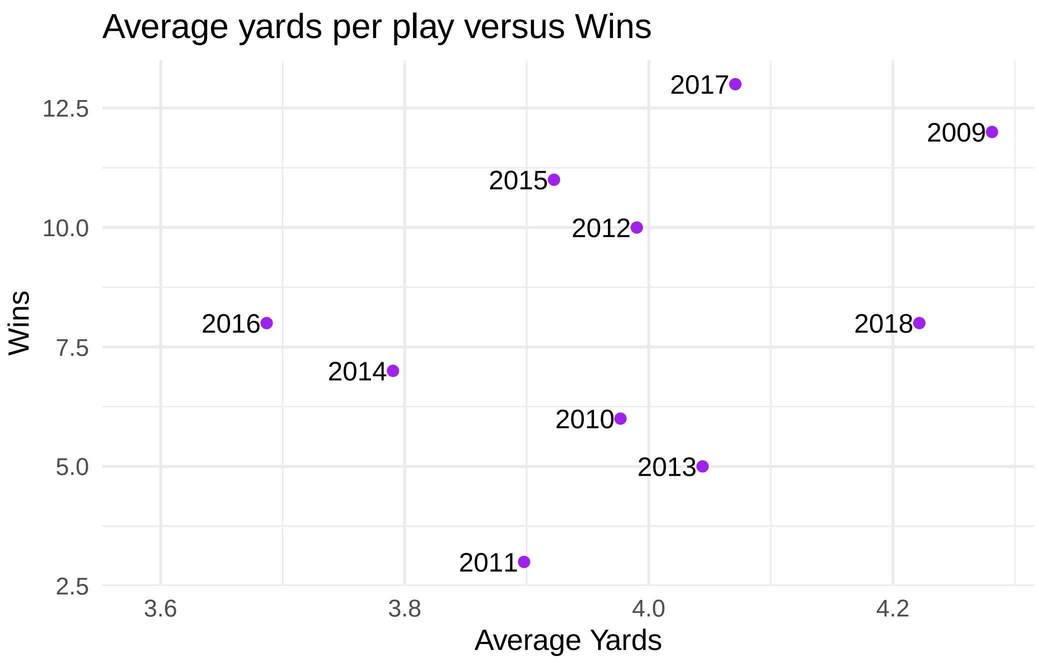
Let’s add the Vikings wins to this table to look at how yards per play relates to games won.

```
# Adding in a vector with wins is easier than trying to extract this information from a pl

wins<-c(12,6,3,10,5,7,11,8,13,8)

vikes_table_2$wins=wins

vikes_table_2
```

```
# A tibble: 10 x 4
    year avg_yards yards_sd  wins
   <dbl>     <dbl>    <dbl> <dbl>
 1  2009      4.28     8.39    12
 2  2010      3.98     7.54     6
 3  2011      3.90     8.14     3
 4  2012      3.99     8.20    10
 5  2013      4.04     8.36     5
 6  2014      3.79     7.70     7
 7  2015      3.92     8.29    11
 8  2016      3.69     7.16     8
 9  2017      4.07     7.83    13
10  2018      4.22     7.83     8
```

Visualization

```
library(ggplot2)
ggplot(vikes_table_2, aes(x=avg_yards, y = wins))+
  geom_point(color="purple")+
  geom_text(aes(label=year), hjust = 1.1, size = 3.5)+
  labs(
    title = "Average yards per play versus Wins",
    x = "Average Yards",
    y = "Wins"
  )+
  theme_minimal()+
   expand_limits(x = min(vikes_table_2$avg_yards) - 0.1)
```



There seems to be a loose positive trend between average yards per play and record.

Let's next look at yards/run plays and yards/pass plays by year.

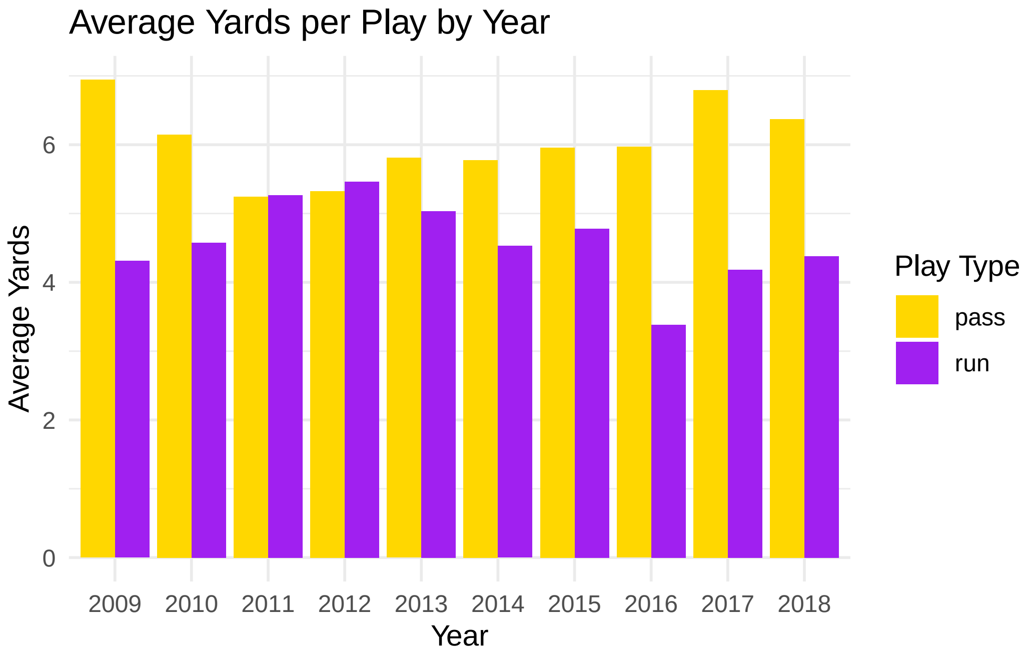
```
vikes_table_3 <- vikes_data |>
  mutate(year = year(game_date)) |>
  group_by(year) |>
  summarize(
    avg_yards = mean(yards_gained, na.rm = TRUE),
    yards_sd  = sd(yards_gained, na.rm = TRUE),
    avg_run_yrds = mean(ifelse(play_type=="run", yards_gained, NA), na.rm=TRUE),
    avg_pass_yrds = mean(ifelse(play_type=="pass", yards_gained, NA), na.rm=TRUE)
  )
vikes_table_3
```

```
# A tibble: 10 x 5
    year avg_yards yards_sd avg_run_yrds avg_pass_yrds
   <dbl>     <dbl>    <dbl>        <dbl>         <dbl>
 1  2009      4.28     8.39         4.31          6.94
 2  2010      3.98     7.54         4.57          6.15
 3  2011      3.90     8.14         5.27          5.25
 4  2012      3.99     8.20         5.46          5.32
 5  2013      4.04    8.36          5.03          5.81
 6  2014      3.79    7.70          4.53          5.78
 7  2015      3.92    8.29          4.78          5.96
 8  2016      3.69    7.16          3.38          5.97
 9  2017      4.07    7.83          4.18          6.80
10  2018      4.22    7.83          4.38          6.37
```

```
# I want to visualize this so I'm going to pivot longer
vikes_table_3_long<-vikes_table_3|>
  rename(
    run = avg_run_yrds,
    pass = avg_pass_yrds
  )|>
  pivot_longer(cols =c(run,pass),
               names_to = "play_type",
               values_to= "average_yards")
vikes_table_3_long
```

```
# A tibble: 20 x 5
    year avg_yards yards_sd play_type average_yards
   <dbl>     <dbl>    <dbl> <chr>             <dbl>
 1  2009      4.28     8.39 run                4.31
 2  2009      4.28     8.39 pass               6.94
 3  2010      3.98     7.54 run                4.57
 4  2010      3.98     7.54 pass               6.15
 5  2011      3.90     8.14 run                5.27
 6  2011      3.90     8.14 pass               5.25
 7  2012      3.99     8.20 run                5.46
 8  2012      3.99     8.20 pass               5.32
 9  2013      4.04     8.36 run                5.03
10  2013      4.04     8.36 pass               5.81
11  2014      3.79     7.70 run                4.53
12  2014      3.79     7.70 pass               5.78
13  2015      3.92     8.29 run                4.78
14  2015      3.92     8.29 pass               5.96
15  2016      3.69     7.16 run                3.38
16  2016      3.69     7.16 pass               5.97
17  2017      4.07     7.83 run                4.18
18  2017      4.07     7.83 pass               6.80
19  2018      4.22     7.83 run                4.38
20  2018      4.22     7.83 pass               6.37
```

```r
ggplot(vikes_table_3_long, aes(x = factor(year), y = average_yards, fill = play_type)) +
  geom_col(position = "dodge") +
  labs(
    title = "Average Yards per Play by Year",
    x = "Year",
    y = "Average Yards",
    fill = "Play Type"
  ) +
    scale_fill_manual(
    values = c("run" = "purple",
               "pass" = "gold")
  ) +
  theme_minimal()
```



- Vikings fans will fondly remember 2012 as Adrian Peterson rushing for 2000+ yards in 2012 after tearing his ACL and winning MVP. SKOL. This year the Vikings had the highest rush yards/attempt of any year.
- Vikings fans will also remember 2009 as the year Brett Favre threw for 4200 yards and took the Vikings to the NFC championship game, only to lose to the Saints after an infamous “too many men on the field” penalty, and “bountygate”--a system that incentivized Saints defensive players to try to knock opposing players out of the game. This was the year with the highest pass yards/attempt.

- Let us also not forget that Favre and Peterson have both endured their fair share of scandals, and I don't want to reminisce on their glory days without noting their complicated legacies.

Lets look at some other things:

```r
vikes_table_4 <- vikes_data |>
  mutate(year = year(game_date)) |>
  group_by(year) |>
  summarize(
    avg_yards = mean(yards_gained, na.rm = TRUE),
    yards_sd  = sd(yards_gained, na.rm = TRUE),
    avg_run_epa = mean(ifelse(play_type=="run", epa, NA), na.rm=TRUE),
    avg_pass_epa = mean(ifelse(play_type=="pass", epa, NA), na.rm=TRUE),
    avg_run_wpa = mean(ifelse(play_type=="run", wpa, NA), na.rm=TRUE),
    avg_pass_wpa = mean(ifelse(play_type=="pass", wpa, NA), na.rm=TRUE)
  )
vikes_table_4
```

```
# A tibble: 10 x 7
    year avg_yards yards_sd avg_run_epa avg_pass_epa avg_run_wpa avg_pass_wpa
   <dbl>     <dbl>    <dbl>       <dbl>        <dbl>       <dbl>        <dbl>
 1  2009      4.28     8.39     -0.153         0.213   -0.00353      0.00506
 2  2010      3.98     7.54     -0.0899       -0.112   -0.00102     -0.00123
 3  2011      3.90     8.14      0.121        -0.128    0.00513     -0.00354
 4  2012      3.99     8.20      0.0183       -0.0450   0.00233     -0.000512
 5  2013      4.04     8.36      0.0405       -0.0944   0.00316     -0.000187
 6  2014      3.79     7.70      0.0171       -0.0520   0.00239     -0.00109
 7  2015      3.92     8.29      0.0219        0.0109  -0.000891     0.00203
 8  2016      3.69     7.16     -0.211         0.0816  -0.00425      0.000850
 9  2017      4.07     7.83     -0.0909        0.182   -0.000631     0.00547
10  2018      4.22     7.83     -0.146         0.0404  -0.00437      0.00274
```

```r
# I want to visualize this so I'm going to pivot longer for EPA
vikes_table_4_long_ep<-vikes_table_4|>
  rename(
    run = avg_run_epa,
    pass = avg_pass_epa
  )|>
  pivot_longer(cols =c(run,pass),
               names_to = "play_type",
               values_to= "average_epa")
```

```r
vikes_table_4_long_ep
```

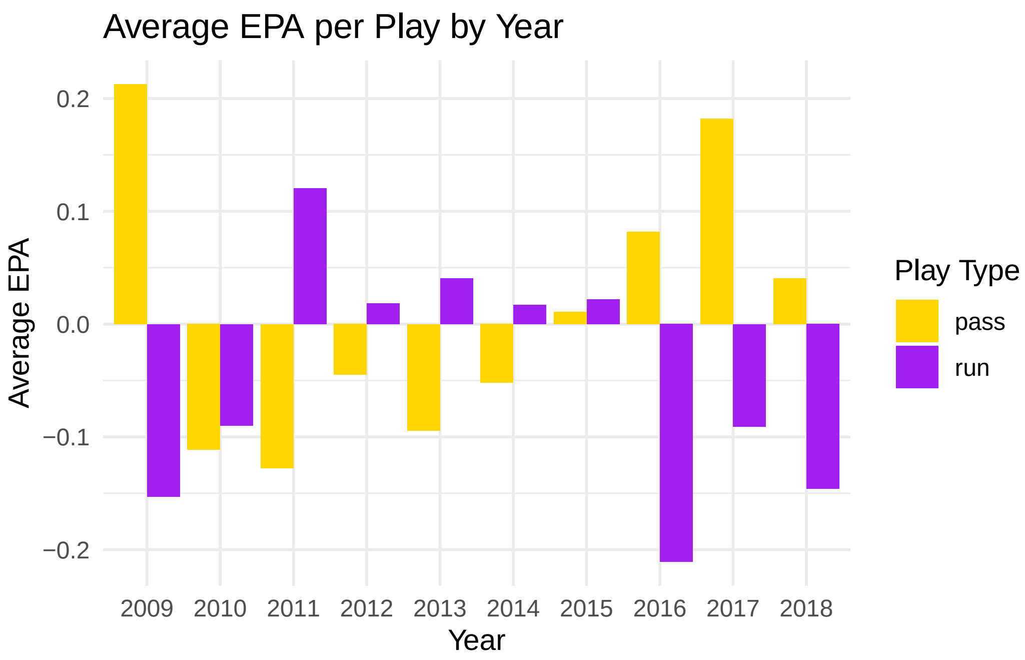
```
# A tibble: 20 x 7
    year avg_yards yards_sd avg_run_wpa avg_pass_wpa play_type average_epa
   <dbl>     <dbl>    <dbl>       <dbl>        <dbl> <chr>           <dbl>
 1  2009      4.28     8.39   -0.00353      0.00506  run           -0.153 
 2  2009      4.28     8.39   -0.00353      0.00506  pass           0.213 
 3  2010      3.98     7.54   -0.00102     -0.00123  run           -0.0899
 4  2010      3.98     7.54   -0.00102     -0.00123  pass          -0.112 
 5  2011      3.90     8.14    0.00513     -0.00354  run            0.121 
 6  2011      3.90     8.14    0.00513     -0.00354  pass          -0.128 
 7  2012      3.99     8.20    0.00233     -0.000512 run            0.0183
 8  2012      3.99     8.20    0.00233     -0.000512 pass          -0.0450
 9  2013      4.04     8.36    0.00316     -0.000187 run            0.0405
10  2013      4.04     8.36    0.00316     -0.000187 pass          -0.0944
11  2014      3.79     7.70    0.00239     -0.00109  run            0.0171
12  2014      3.79     7.70    0.00239     -0.00109  pass          -0.0520
13  2015      3.92     8.29   -0.000891     0.00203  run            0.0219
14  2015      3.92     8.29   -0.000891     0.00203  pass           0.0109
15  2016      3.69     7.16   -0.00425      0.000850 run           -0.211 
16  2016      3.69     7.16   -0.00425      0.000850 pass           0.0816
17  2017      4.07     7.83   -0.000631     0.00547  run           -0.0909
18  2017      4.07     7.83   -0.000631     0.00547  pass           0.182 
19  2018      4.22     7.83   -0.00437      0.00274  run           -0.146 
20  2018      4.22     7.83   -0.00437      0.00274  pass           0.0404
```

```r
## And pivot longer for WPA
vikes_table_4_long_wp<-vikes_table_4|>
  rename(
    run = avg_run_wpa,
    pass = avg_pass_wpa
  )|>
  pivot_longer(cols =c(run,pass),
               names_to = "play_type",
               values_to= "average_wpa")
vikes_table_4_long_wp
```

```
# A tibble: 20 x 7
    year avg_yards yards_sd avg_run_epa avg_pass_epa play_type average_wpa
   <dbl>     <dbl>    <dbl>       <dbl>        <dbl> <chr>           <dbl>
```

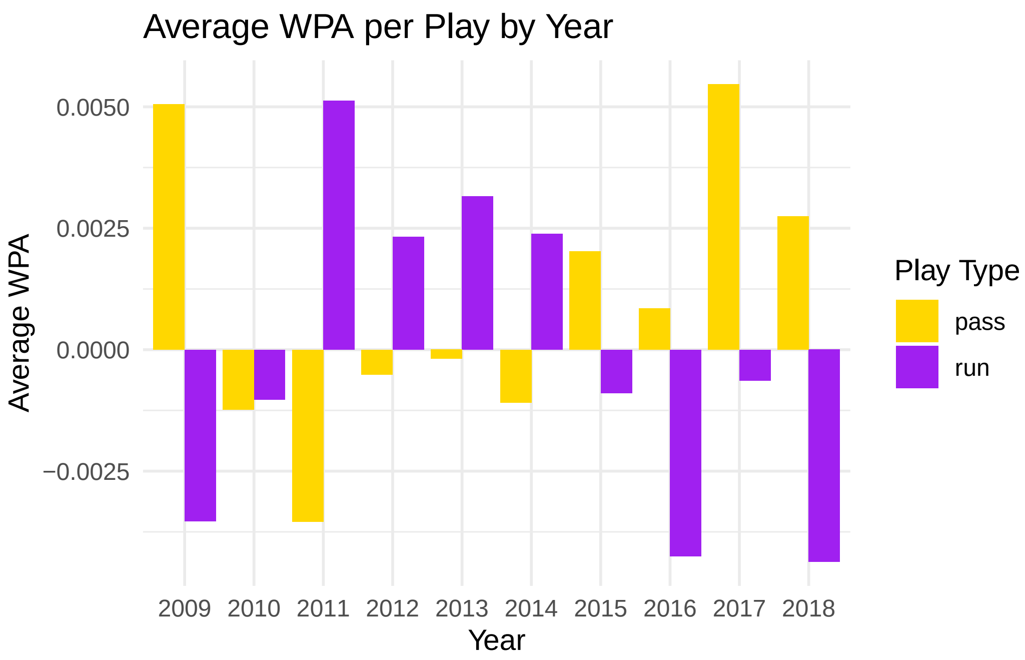
```
 1  2009     4.28     8.39     -0.153        0.213  run          -0.00353
 2  2009     4.28     8.39     -0.153        0.213  pass          0.00506
 3  2010     3.98     7.54     -0.0899      -0.112  run          -0.00102
 4  2010     3.98     7.54     -0.0899      -0.112  pass         -0.00123
 5  2011     3.90     8.14      0.121       -0.128  run           0.00513
 6  2011     3.90     8.14      0.121       -0.128  pass         -0.00354
 7  2012     3.99     8.20      0.0183      -0.0450 run           0.00233
 8  2012     3.99     8.20      0.0183      -0.0450 pass         -0.000512
 9  2013     4.04     8.36      0.0405      -0.0944 run           0.00316
10  2013     4.04     8.36      0.0405      -0.0944 pass         -0.000187
11  2014     3.79     7.70      0.0171      -0.0520 run           0.00239
12  2014     3.79     7.70      0.0171      -0.0520 pass         -0.00109
13  2015     3.92     8.29      0.0219       0.0109 run          -0.000891
14  2015     3.92     8.29      0.0219       0.0109 pass          0.00203
15  2016     3.69     7.16     -0.211        0.0816 run          -0.00425
16  2016     3.69     7.16     -0.211        0.0816 pass          0.000850
17  2017     4.07     7.83     -0.0909       0.182  run          -0.000631
18  2017     4.07     7.83     -0.0909       0.182  pass          0.00547
19  2018     4.22     7.83     -0.146        0.0404 run          -0.00437
20  2018     4.22     7.83     -0.146        0.0404 pass          0.00274
```

```r
ggplot(vikes_table_4_long_ep, aes(x = factor(year), y = average_epa, fill = play_type)) +
  geom_col(position = "dodge") +
  labs(
    title = "Average EPA per Play by Year",
    x = "Year",
    y = "Average EPA",
    fill = "Play Type"
  ) +
    scale_fill_manual(
    values = c("run" = "purple",
               "pass" = "gold")
  ) +
  theme_minimal()
```



- Similar trends are visible here. Note that the 2010 Vikings were 6-10.

```
ggplot(vikes_table_4_long_wp, aes(x = factor(year), y = average_wpa, fill = play_type)) +
  geom_col(position = "dodge") +
  labs(
    title = "Average WPA per Play by Year",
    x = "Year",
    y = "Average WPA",
    fill = "Play Type"
  ) +
    scale_fill_manual(
    values = c("run" = "purple",
               "pass" = "gold")
  ) +
  theme_minimal()
```



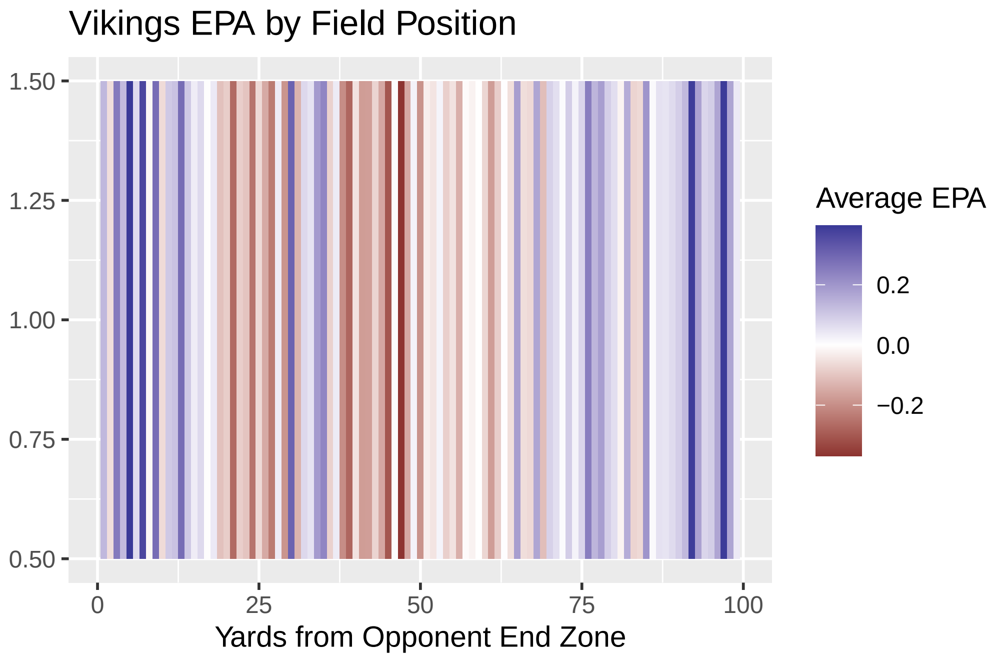
- This is also a fascinating breakdown. Note that in 2009, running was not advantageous at all, although Adrian Peterson did run for 1300+ yards that year.

Let's try to visualize some other things. I want to try do a heatmap relating EPA and field position.

```
epa_by_field<-vikes_data|>
  group_by(yardline_100)|>
  summarize(avg_epa = mean(epa, na.rm=TRUE))
epa_by_field
```

```
# A tibble: 99 x 2
   yardline_100 avg_epa
          <dbl>   <dbl>
 1            1  0.133
 2            2 -0.0574
 3            3  0.256
 4            4  0.129
 5            5  0.396
 6            6  0.0666
 7            7  0.367
 8            8 -0.0145
 9            9  0.284
10           10 -0.0633
# i 89 more rows
```

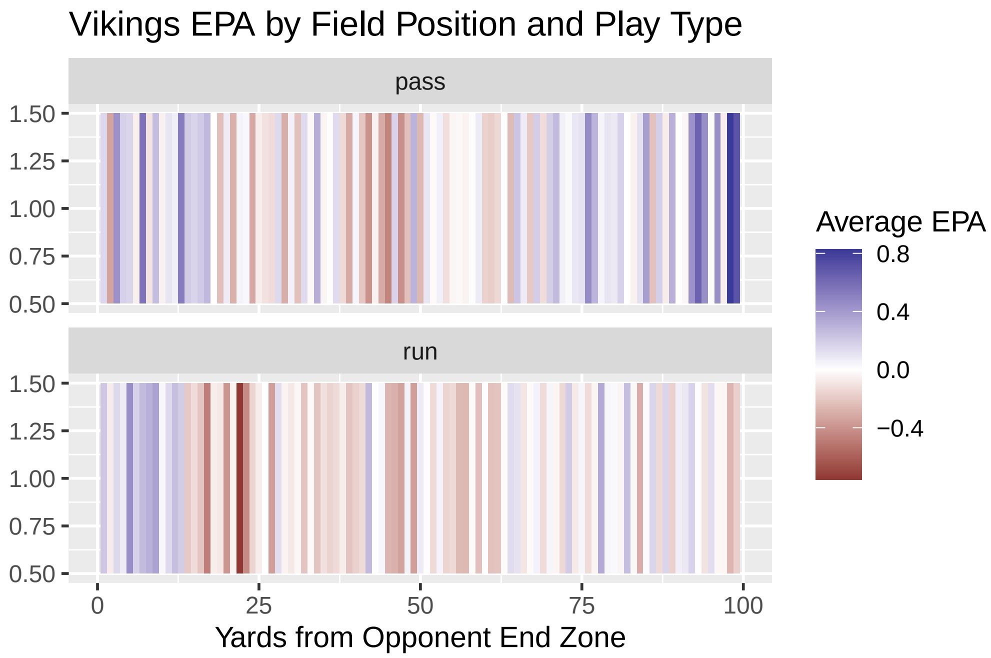
```
ggplot(epa_by_field, aes(x=yardline_100, y=1, fill = avg_epa))+
  geom_tile()+
  scale_fill_gradient2()+
  labs(
    title = "Vikings EPA by Field Position",
    x = "Yards from Opponent End Zone",
    y = "",
    fill = "Average EPA"
  )
```



I would like to facet this by play type.

```
epa_by_field_type<-vikes_data|>
  filter(play_type %in% c("run","pass"))|>
  group_by(yardline_100,play_type)|>
  summarize(avg_epa = mean(epa, na.rm = TRUE), .groups = "drop")
```

```r
ggplot(epa_by_field_type, aes(x=yardline_100, y=1, fill=avg_epa))+
       geom_tile()+
       scale_fill_gradient2()+
       facet_wrap(~play_type, ncol = 1) +
       labs(
    title = "Vikings EPA by Field Position and Play Type",
    x = "Yards from Opponent End Zone",
    y = "",
    fill = "Average EPA"
  )
```



Seems like passing generally has a higher EPA from just about anywhere in the field.