Vikings_EDA

Exploring Vikings NFL Data

```
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
```

```
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.
```

Contingency Tables and Numeric Summaries

First, visualize vikings play type per down.

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

```
    extra_point field_goal kickoff no_play pass punt qb_kneel qb_spike  run
  1           0         17       0     516 4169    0      133       35 4522
  2           0         17       0     400 3800    0       58        1 2771
  3           0         16       0     371 3304    0       34        2  818
  4           0        562       0     118  194 1488        1        0  102
```

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.

```
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       758       1098     0.408      0.592
 2  2010       832       1088     0.433      0.567
 3  2011       872       1142     0.433      0.567
 4  2012       942       1246     0.431      0.569
 5  2013       843       1275     0.398      0.602
 6  2014       837       1136     0.424      0.576
 7  2015       811       1028     0.441      0.559
 8  2016       758       1228     0.382      0.618
 9  2017       897       1198     0.428      0.572
10  2018       678       1057     0.391      0.609
```

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      3.94     8.08
 2  2010      3.77     7.52
 3  2011      3.94     7.95
 4  2012      3.79     7.64
 5  2013      4.11     8.31
 6  2014      3.79     7.47
 7  2015      3.86     8.06
 8  2016      3.60     7.23
 9  2017      3.66     7.36
10  2018      3.87     7.79
```

Cool! Vikings were averaging a high 4.1 yards per play in 2013. Surprisingly, the Vikings were 5-10-1 in spite of this.

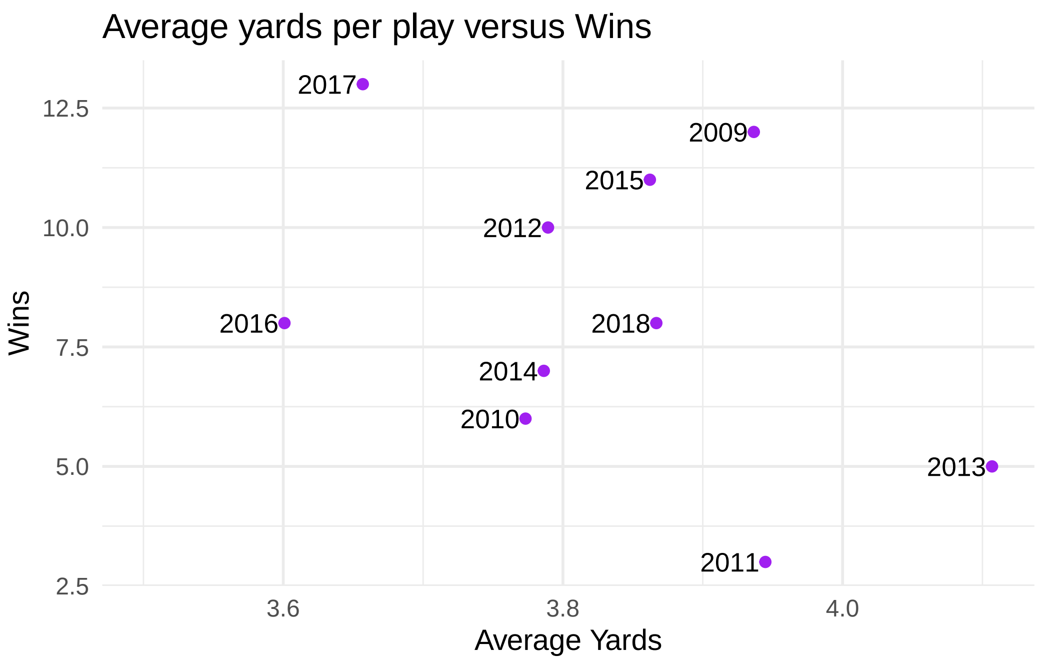
Let's add the Vikings wins to this table to look at how yards per play relates to games won.

```
# 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      3.94     8.08    12
 2  2010          3.77     7.52      6
 3  2011          3.94     7.95      3
 4  2012          3.79     7.64     10
 5  2013          4.11     8.31      5
 6  2014          3.79     7.47      7
 7  2015          3.86     8.06     11
 8  2016          3.60     7.23      8
 9  2017          3.66     7.36     13
10  2018          3.87     7.79      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 is no obvious relationship between these two variables.

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

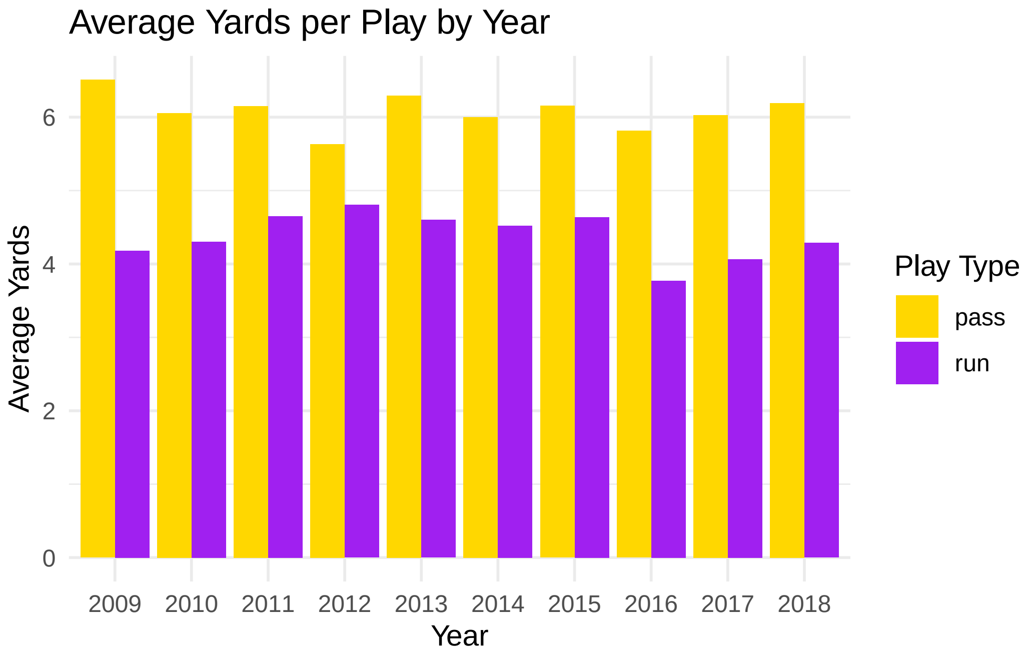
```
vikes_table_3 <- vikes_data |>
  mutate(year = year(game_date)) |>
  group_by(year) |>
  summarize(
    avg_yards = mean(yards_gained, na.rm = TRUE),
    yards_sd  = sd(yards_gained, na.rm = TRUE),
    avg_run_yrds = mean(ifelse(play_type=="run", yards_gained, NA), na.rm=TRUE),
    avg_pass_yrds = 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      3.94     8.08         4.18          6.51
 2  2010      3.77     7.52         4.31          6.06
 3  2011      3.94     7.95         4.65          6.15
 4  2012      3.79     7.64         4.81          5.63
 5  2013      4.11     8.31          4.60          6.29
 6  2014      3.79     7.47          4.52          6.00
 7  2015      3.86     8.06          4.64          6.15
 8  2016      3.60     7.23          3.77          5.81
 9  2017      3.66     7.36          4.06          6.03
10  2018      3.87     7.79          4.29          6.19
```

```
# 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      3.94     8.08 run                4.18
 2  2009      3.94     8.08 pass               6.51
 3  2010      3.77     7.52 run                4.31
 4  2010      3.77     7.52 pass               6.06
 5  2011      3.94     7.95 run                4.65
 6  2011      3.94     7.95 pass               6.15
 7  2012      3.79     7.64 run                4.81
 8  2012      3.79     7.64 pass               5.63
 9  2013      4.11     8.31 run                4.60
10  2013      4.11     8.31 pass               6.29
11  2014      3.79     7.47 run                4.52
12  2014      3.79     7.47 pass               6.00
13  2015      3.86     8.06 run                4.64
14  2015      3.86     8.06 pass               6.15
15  2016      3.60     7.23 run                3.77
16  2016      3.60     7.23 pass               5.81
17  2017      3.66     7.36 run                4.06
18  2017      3.66     7.36 pass               6.03
19  2018      3.87     7.79 run                4.29
20  2018      3.87     7.79 pass               6.19
```

```
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 this.

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      3.94     8.08    -0.168        0.118     -0.00372      0.00282
 2  2010      3.77     7.52    -0.0978      -0.0806    -0.000535    -0.00118
 3  2011      3.94     7.95     0.0108       0.0373     0.00229      0.00244
 4  2012      3.79     7.64    -0.0451       0.0145     0.000343     0.00138
 5  2013      4.11     8.31     0.00710      0.0360     0.000662     0.00244
 6  2014      3.79     7.47     0.00778      0.00918    0.00195      0.00178
 7  2015      3.86     8.06    -0.0299       0.0298    -0.00137      0.00150
 8  2016      3.60     7.23    -0.152        0.0166    -0.00288     -0.000410
 9  2017      3.66     7.36    -0.112        0.0315    -0.00160      0.00127
10  2018      3.87     7.79    -0.0822      -0.0276    -0.00199      0.00158
```

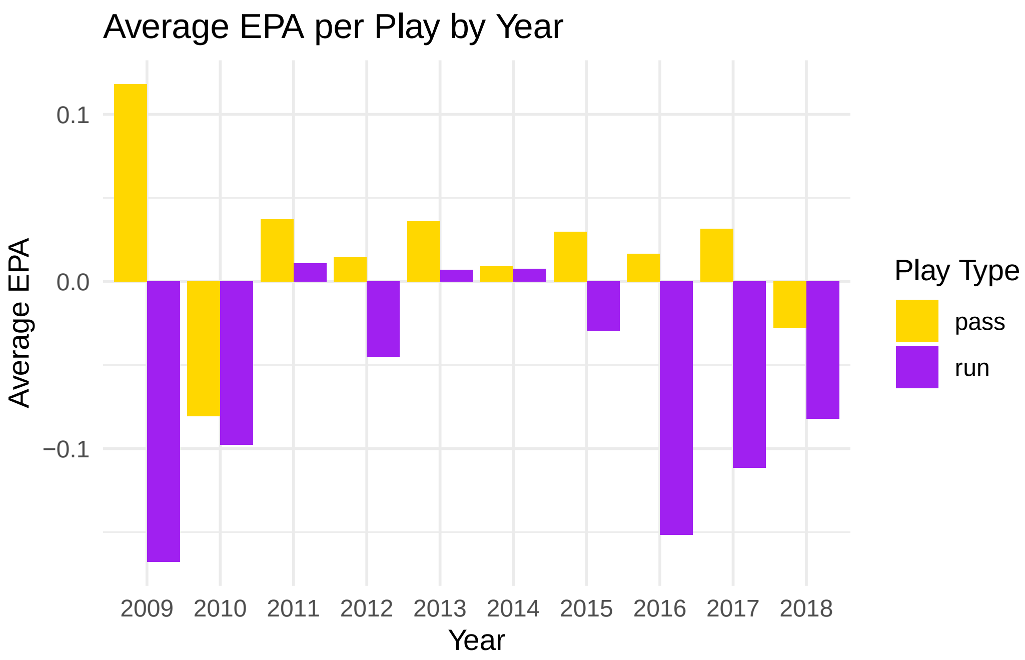
```r
# I want to visualize this so I'm going to pivot longer for EPA
vikes_table_4_long_ep<-vikes_table_4|>
  rename(
    run = avg_run_epa,
    pass = avg_pass_epa
  )|>
  pivot_longer(cols =c(run,pass),
               names_to = "play_type",
               values_to= "average_epa")
vikes_table_4_long_ep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

```
# 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      3.94     8.08   -0.00372      0.00282  run           -0.168
 2  2009      3.94     8.08   -0.00372      0.00282  pass           0.118
 3  2010      3.77     7.52   -0.000535    -0.00118  run           -0.0978
 4  2010      3.77     7.52   -0.000535    -0.00118  pass          -0.0806
 5  2011      3.94     7.95    0.00229      0.00244  run            0.0108
 6  2011      3.94     7.95    0.00229      0.00244  pass           0.0373
 7  2012      3.79     7.64    0.000343     0.00138  run           -0.0451
 8  2012      3.79     7.64    0.000343     0.00138  pass           0.0145
 9  2013      4.11     8.31    0.000662     0.00244  run            0.00710
10  2013      4.11     8.31    0.000662     0.00244  pass           0.0360
11  2014      3.79     7.47    0.00195      0.00178  run            0.00778
12  2014      3.79     7.47    0.00195      0.00178  pass           0.00918
13  2015      3.86     8.06   -0.00137      0.00150  run           -0.0299
14  2015      3.86     8.06   -0.00137      0.00150  pass           0.0298
15  2016      3.60     7.23   -0.00288     -0.000410 run           -0.152
16  2016      3.60     7.23   -0.00288     -0.000410 pass           0.0166
17  2017      3.66     7.36   -0.00160      0.00127  run           -0.112
18  2017      3.66     7.36   -0.00160      0.00127  pass           0.0315
19  2018      3.87     7.79   -0.00199      0.00158  run           -0.0822
20  2018      3.87     7.79   -0.00199      0.00158  pass          -0.0276
```

```
## 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
```

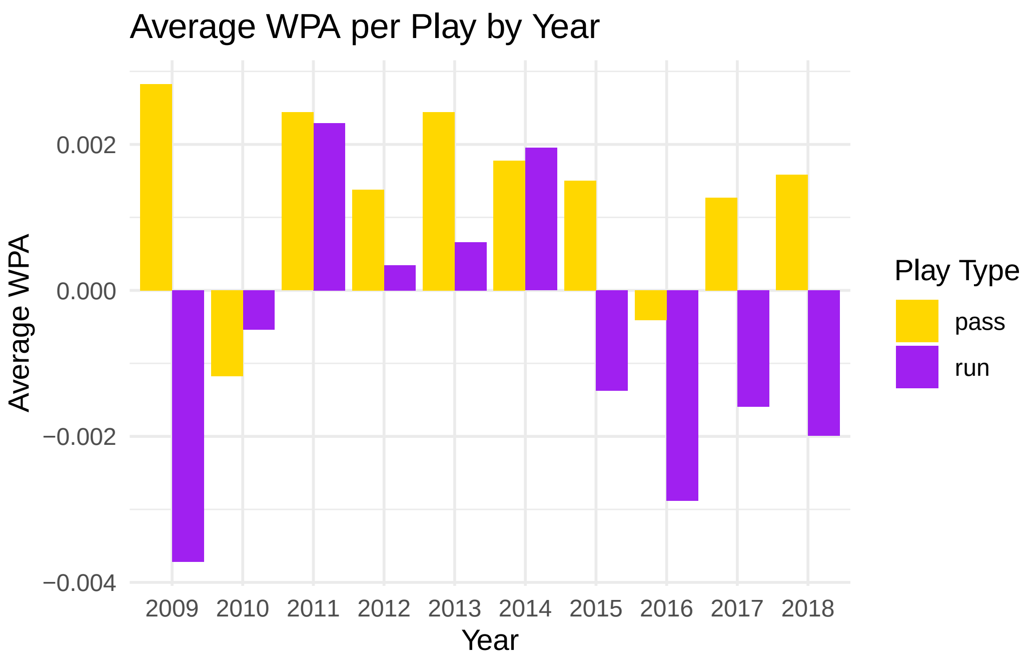
```
# A tibble: 20 x 7
    year avg_yards yards_sd avg_run_epa avg_pass_epa play_type average_wpa
   <dbl>     <dbl>    <dbl>       <dbl>        <dbl> <chr>           <dbl>
 1  2009      3.94     8.08    -0.168        0.118   run         -0.00372
 2  2009      3.94     8.08    -0.168        0.118   pass         0.00282
 3  2010      3.77     7.52   -0.0978      -0.0806   run         -0.000535
 4 2010      3.77     7.52    -0.0978      -0.0806  pass         -0.00118
 5 2011      3.94     7.95     0.0108       0.0373  run           0.00229
 6 2011      3.94     7.95     0.0108       0.0373  pass          0.00244
 7 2012      3.79     7.64    -0.0451       0.0145  run           0.000343
 8 2012      3.79     7.64    -0.0451       0.0145  pass          0.00138
 9 2013      4.11     8.31     0.00710      0.0360  run           0.000662
10 2013      4.11     8.31     0.00710      0.0360  pass          0.00244
11 2014      3.79     7.47     0.00778      0.00918 run           0.00195
12 2014      3.79     7.47     0.00778      0.00918 pass          0.00178
13 2015      3.86     8.06    -0.0299       0.0298  run          -0.00137
14 2015      3.86     8.06    -0.0299       0.0298  pass          0.00150
15 2016      3.60     7.23    -0.152        0.0166  run          -0.00288
16 2016      3.60     7.23    -0.152        0.0166  pass         -0.000410
17 2017      3.66     7.36    -0.112        0.0315  run          -0.00160
18 2017      3.66     7.36    -0.112        0.0315  pass          0.00127
19 2018      3.87     7.79    -0.0822      -0.0276  run          -0.00199
20 2018      3.87     7.79    -0.0822      -0.0276  pass          0.00158
```

```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.

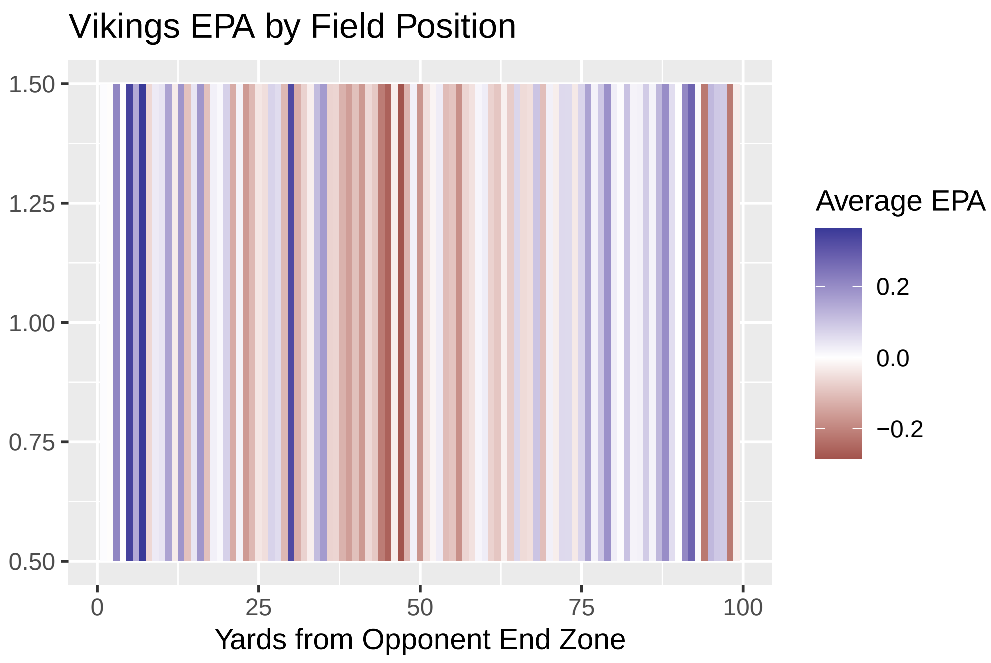
```
epa_by_field<-vikes_data|>
  group_by(yardline_100)|>
  summarize(avg_epa = mean(epa, na.rm=TRUE))
epa_by_field
```

```
# A tibble: 100 x 2
   yardline_100  avg_epa
          <dbl>    <dbl>
 1            1  0.00506
 2            2 -0.00389
 3            3  0.211
 4            4 -0.00103
 5            5  0.344
 6            6  0.146
 7            7  0.362
 8            8 -0.0618
 9            9  0.0346
```

```
10            10  0.0460
# i 90 more rows
```

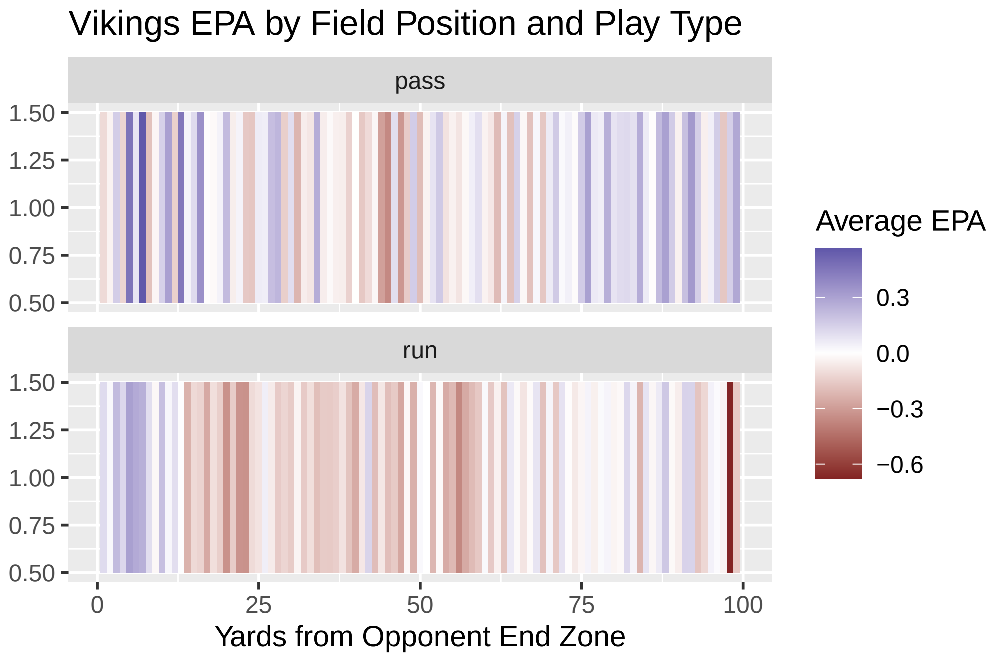
```
ggplot(epa_by_field, aes(x=yardline_100, y=1, fill = avg_epa))+
  geom_tile()+
  scale_fill_gradient2()+
  labs(
    title = "Vikings EPA by Field Position",
    x = "Yards from Opponent End Zone",
    y = "",
    fill = "Average EPA"
  )
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_tile()`).
```



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")

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.