Big Data Bowl 2023

Sam Burch

Introduction

The theme of this year's Big Data Bowl is about researching pass blocking and pass rushing in the NFL. Specifically, we will go through sacks on both sides of the ball. Sacks are where a quarterback is tackled behind the line of scrimmage. This can be very detrimental for the offense, even as bad as a turnover. According to the Washington Post, a sack (on average) can be worth about 1.75 point to the defense. So, we seek to find out what affects sacks and which players are best at limiting these on offense and maximizing these on defense. (Note: It has been proven that quarterbacks largely are responsible for sacks. Per the Big Lead, a QBs Sack Rate is very stable when changing teams. However, because of the impact of a sack and other positions being liable as well, this will still be our base measure of play.)

We will be using the data provided for this competition. Here, we import such data (on top of the libraries we will be using).

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggrepel)


games = read.csv('games.csv', header = TRUE)
pff = read.csv('pffScoutingData.csv', header = TRUE)
players = read.csv('players.csv', header = TRUE)
plays = read.csv('plays.csv', header=TRUE)

games = as_tibble(games)
pff = as_tibble(pff)
players = as_tibble(players)
plays = as_tibble(plays)
```

Offense

Total Sacks and Pressures Allowed

Let us consider the total sacks allowed by position.

```r
pff |>
  filter(pff_role == 'Pass Block', pff_sackAllowed == 1) |>
  group_by(pff_positionLinedUp) |>
  summarise(sacks = n()) |>
  arrange(-sacks)
```

```
## # A tibble: 15 x 2
##    pff_positionLinedUp sacks
##    <chr>               <int>
##  1 RT                     88
##  2 LT                     79
##  3 RG                     50
##  4 LG                     48
##  5 C                      40
##  6 HB-R                   13
##  7 HB-L                   12
##  8 TE-L                   10
##  9 HB                      9
## 10 TE-R                    8
## 11 TE-iR                   4
## 12 TE-iL                   3
## 13 FB                      1
## 14 SLWR                    1
## 15 TE-oR                   1
```

The offensive line players – tackles, guards, and centers – have by far the highest number of sacks allowed. This is likely due to the high number of pass blocking snaps they have compared to other positions. So, we need to find a way to combat this issue. Before we do so, let us look at another statistic.

Pressures (defined as a hit, hurry, or sack on the quarterback) have been proven to be more stable than sacks and are even quite predictive of future sacks (per PFF). So, let us consider the total pressures allowed by position – min 10 pressures.

```r
pff |>
  filter(pff_role == 'Pass Block',
         pff_hurryAllowed == 1 |
         pff_sackAllowed == 1  |
         pff_hitAllowed == 1) |>
  group_by(pff_positionLinedUp) |>
  summarise(pressures = n()) |>
  arrange(-pressures) |>
  filter(pressures > 10)
```

```
## # A tibble: 12 x 2
##    pff_positionLinedUp pressures
##    <chr>                   <int>
##  1 RT                        660
##  2 LT                        621
##  3 RG                        552
##  4 LG                        479
##  5 C                         354
##  6 HB-L                       75
##  7 HB-R                       68
```

```
##  8 TE-R                        46
##  9 HB                          34
## 10 TE-L                        27
## 11 TE-iR                       16
## 12 TE-oR                       12
```

With pressures allowed, we run into the same issue. This is because total sacks and pressures allowed are volume statistics. Just like passing yards, the more pass attempts you have, the more passing yards you tend to have. Thus, there exists passing yards per attempt. We will do the same here by creating rate statistics.

Offensive Sack and Pressure Rates

To make an offensive sack rate, we will group by position and divide sacks allowed by the number of pass blocking reps. This is what we end up with.

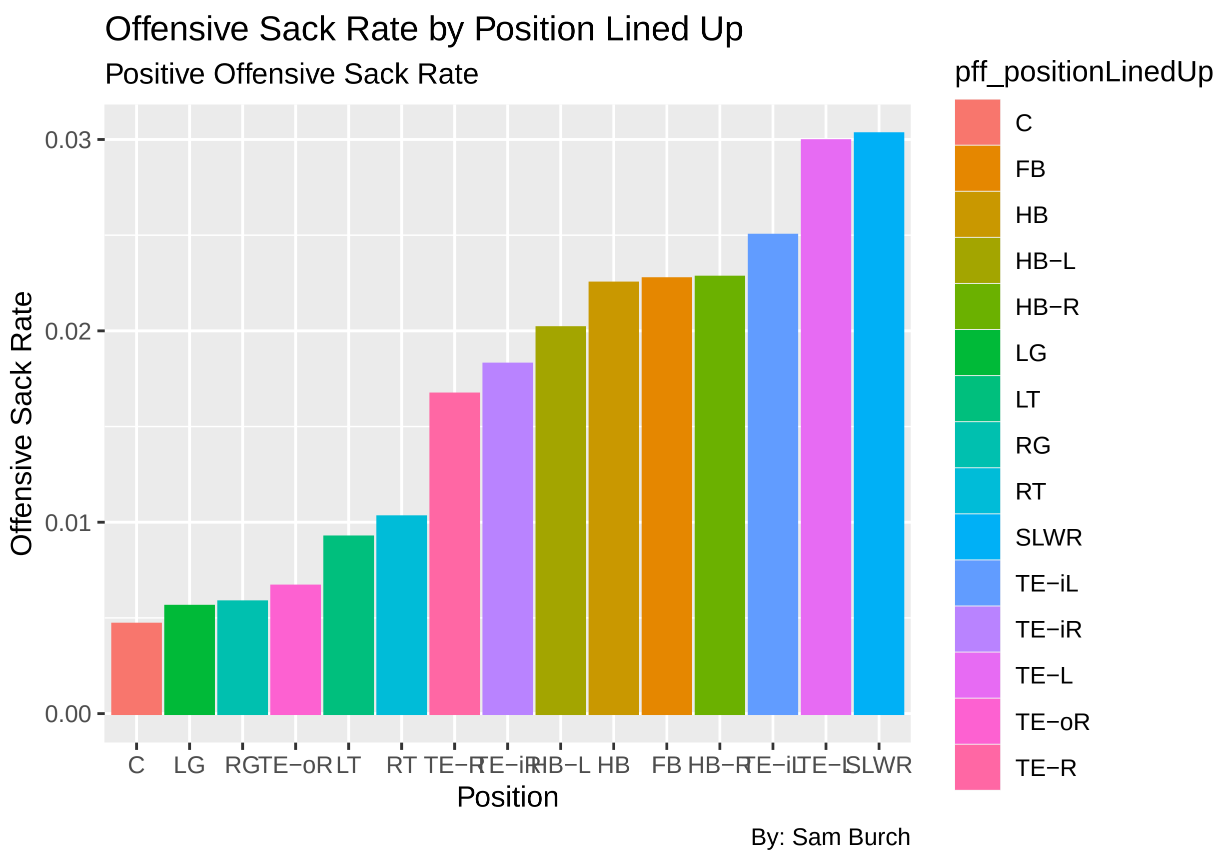
```
sack_rate_plus = pff |>
  filter(pff_role == 'Pass Block') |>
  group_by(pff_positionLinedUp) |>
  summarise(sack_rate = mean(pff_sackAllowed), sacks = sum(pff_sackAllowed), plays = n()) |>
  arrange(-sack_rate) |>
  filter(sack_rate > 0)
sack_rate_plus
```

```
## # A tibble: 15 x 4
##    pff_positionLinedUp sack_rate sacks plays
##    <chr>                   <dbl> <int> <int>
##  1 SLWR                  0.0303      1    33
##  2 TE-L                  0.0299     10   334
##  3 TE-iL                 0.025       3   120
##  4 HB-R                  0.0228     13   570
##  5 FB                    0.0227      1    44
##  6 HB                    0.0225      9   400
##  7 HB-L                  0.0202     12   595
##  8 TE-iR                 0.0183      4   219
##  9 TE-R                  0.0167      8   479
## 10 RT                    0.0103     88  8555
## 11 LT                    0.00923    79  8556
## 12 TE-oR                 0.00667     1   150
## 13 RG                    0.00584    50  8557
## 14 LG                    0.00561    48  8557
## 15 C                     0.00467    40  8557
```

Here is a visual of what the table says about offensive sack rate. Note: Keep in mind the sample size per position.

```
ggplot(sack_rate_plus, aes(reorder(pff_positionLinedUp, sack_rate), sack_rate)) +
  geom_col(aes(color = pff_positionLinedUp, fill = pff_positionLinedUp)) +
  # theme(axis.text.x = element_blank()) +
  labs(
    title = 'Offensive Sack Rate by Position Lined Up',
    subtitle = 'Positive Offensive Sack Rate',
    caption = 'By: Sam Burch',
```

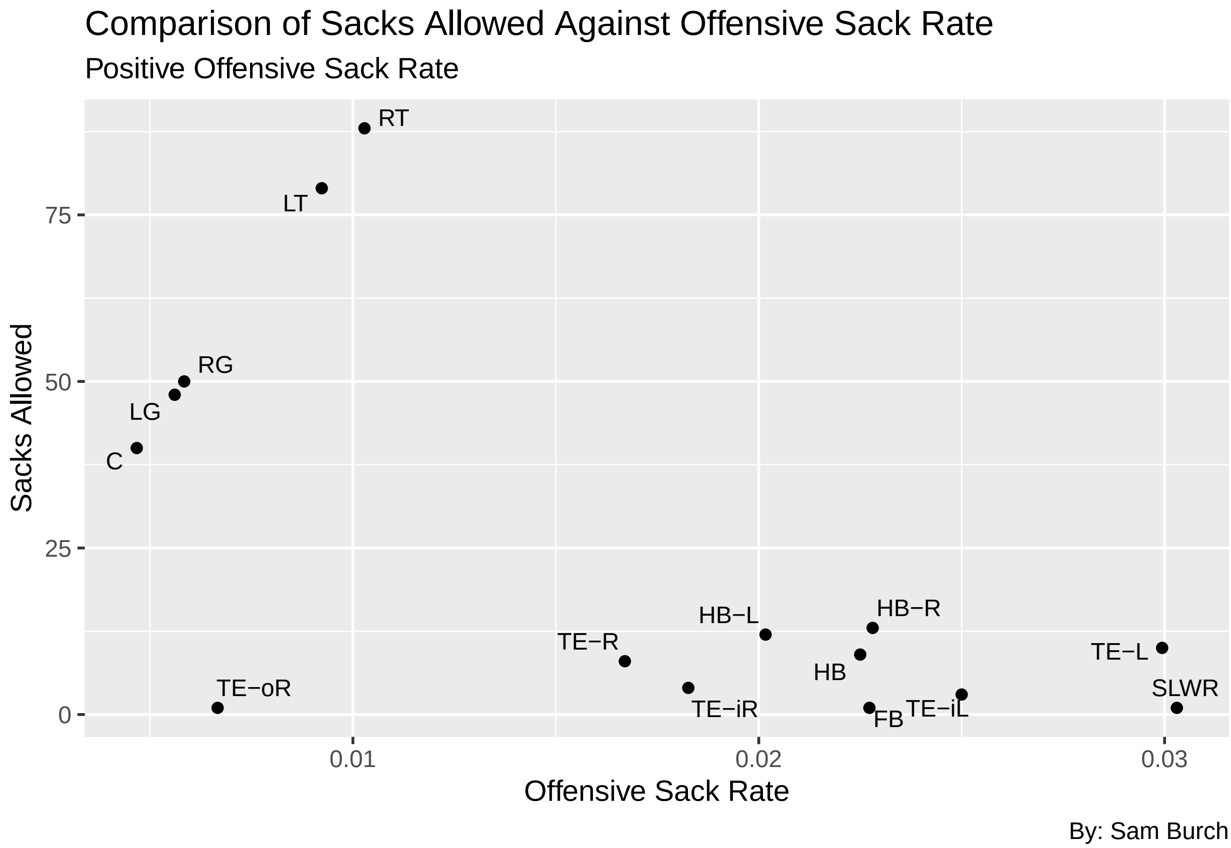
```
    y = 'Offensive Sack Rate',
    x = 'Position'
  )
```



Unlike total sacks allowed, non o-lineman have the higher number here. There are a few possible reasons for this; these positions not being as good of blockers, or even them being the last line of defense for the quarterback. Also, we must take into account that with an increase in volume, it is harder to obtain a high efficiency. However, since having a lower rate here is better, this “helps” the o-lineman.

Let us compare our rate and volume metric now.

```
ggplot(sack_rate_plus, aes(sack_rate, sacks)) +
  geom_point() +
  geom_text_repel(aes(label = pff_positionLinedUp), size = 3) +
  labs(
    title = 'Comparison of Sacks Allowed Against Offensive Sack Rate',
    subtitle = 'Positive Offensive Sack Rate',
    caption = 'By: Sam Burch',
    y = 'Sacks Allowed',
    x = 'Offensive Sack Rate'
  )
```



The first observation from this chart are the two clusters – o-lineman (cluster 1) and non o-lineman (cluster 2). With cluster 1, it is easy to see the higher sacks allowed and lower offensive sack rate. One observation we didn't realize earlier was the positive correlation. This shows how interior o-lineman generate a lower amount of sacks allowed and a lower offensive sack rate compared to the tackles. This makes sense as pass rushing is generally easier to win on the outside. On top of that, we can see players on the right are generally worse than on the left. This might be because teams still prioritize protecting the blindside of the quarterback.

Cluster 2 shows a lack of correlation. Not much can be taken away from this, but maybe as players at other positions give up more sacks, coach's will make sure these players do not keep receiving more snaps. That's why the players with a lower sack rate here have given up a similar amount of sacks as those with a higher sack rate.

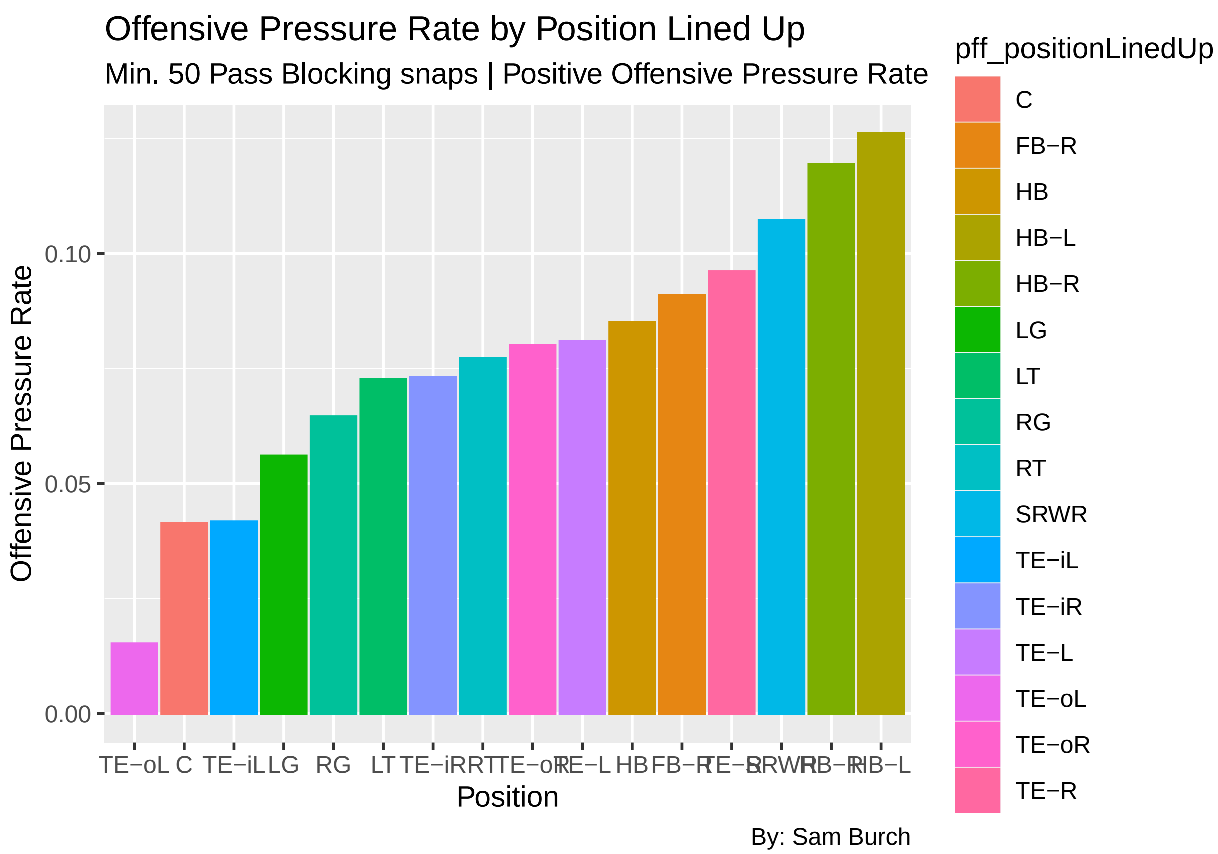
Now we add in offensive pressure rate.

```
pressure_rate_plus = pff |>
  filter(pff_role == 'Pass Block') |>
  mutate(pressureAllowed = case_when((pff_hurryAllowed == 1 |
                                      pff_sackAllowed == 1  |
                                      pff_hitAllowed == 1) ~ 1,
                                     (pff_hurryAllowed == 0 &
                                      pff_sackAllowed == 0  &
                                      pff_hitAllowed == 0) ~ 0)) |>
  group_by(pff_positionLinedUp) |>
  summarise(pressure_rate = mean(pressureAllowed), pressures = sum(pressureAllowed), plays = n()) |>
  arrange(-pressure_rate) |>
  filter(pressure_rate > 0,
         plays >= 50)
```

```r
pressure_rate_plus
```

```
## # A tibble: 16 x 4
##    pff_positionLinedUp pressure_rate pressures plays
##    <chr>                       <dbl>     <dbl> <int>
##  1 HB-L                       0.126         75   595
##  2 HB-R                       0.119         68   570
##  3 SRWR                       0.107          6    56
##  4 TE-R                       0.0960        46   479
##  5 FB-R                       0.0909         5    55
##  6 HB                         0.085         34   400
##  7 TE-L                       0.0808        27   334
##  8 TE-oR                      0.08          12   150
##  9 RT                         0.0771       660  8555
## 10 TE-iR                      0.0731        16   219
## 11 LT                         0.0726       621  8556
## 12 RG                         0.0645       552  8557
## 13 LG                         0.0560       479  8557
## 14 TE-iL                      0.0417         5   120
## 15 C                          0.0414       354  8557
## 16 TE-oL                      0.0152         1    66
```

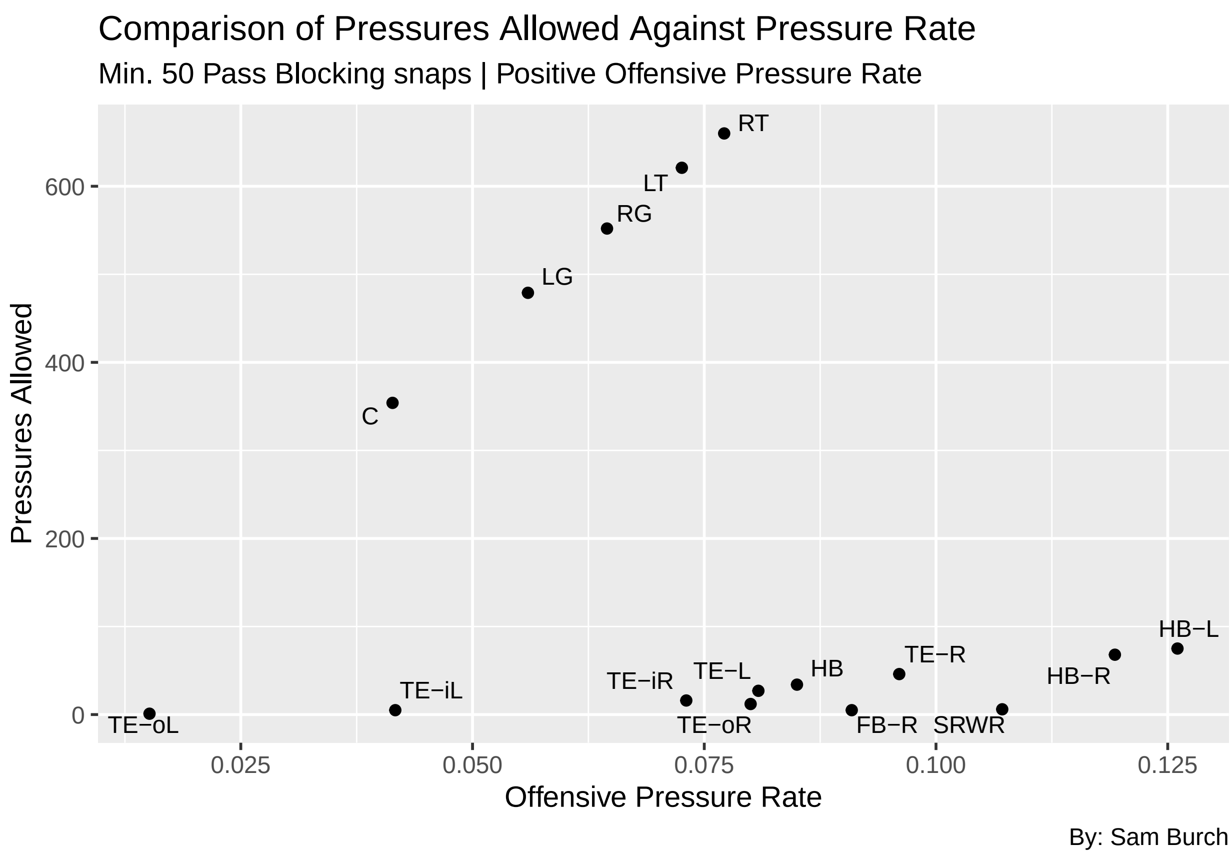
```r
ggplot(pressure_rate_plus, aes(reorder(pff_positionLinedUp, pressure_rate), pressure_rate)) +
  geom_col(aes(color = pff_positionLinedUp, fill = pff_positionLinedUp)) +
  # theme(axis.text.x = element_blank()) +
  labs(
    title = 'Offensive Pressure Rate by Position Lined Up',
    subtitle = 'Min. 50 Pass Blocking snaps | Positive Offensive Pressure Rate',
    caption = 'By: Sam Burch',
    y = 'Offensive Pressure Rate',
    x = 'Position'
  )
```



Besides tight ends showing up slightly better here, this is similar to what we saw with offensive sack rate. Also, keep in mind we added in a play minimum above to eliminate some noise.

Another comparison of metrics.

```
ggplot(pressure_rate_plus, aes(pressure_rate, pressures)) +
  geom_point() +
  geom_text_repel(aes(label = pff_positionLinedUp), size = 3) +
  labs(
    title = 'Comparison of Pressures Allowed Against Pressure Rate',
    subtitle = 'Min. 50 Pass Blocking snaps | Positive Offensive Pressure Rate',
    caption = 'By: Sam Burch',
    y = 'Pressures Allowed',
    x = 'Offensive Pressure Rate'
  )
```



The two clusters appear again. The main difference again is the TE performance. This may be because of o-line players lining up as TEs a variety of times in max protection scenarios. Another reason could simply be TEs blocking weak rushers as opposed to tackles blocking edge defenders.

Offensive Sack Rate vs Offensive Pressure Rate

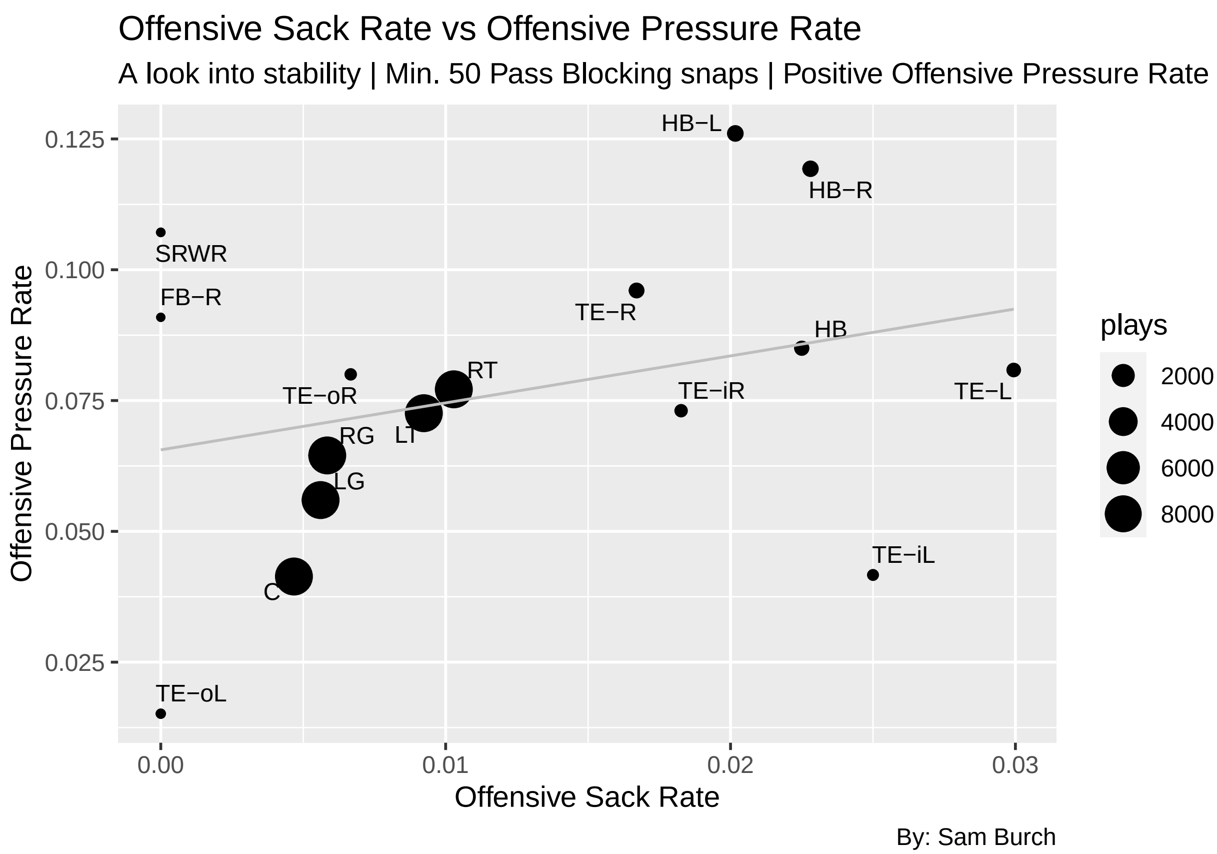
We already know that pressures are a very good measure of future sack performance. So, let us compare the rates of both.

```
sr_pr = pff |>
  filter(pff_role == 'Pass Block') |>
  mutate(pressureAllowed = case_when((pff_hurryAllowed == 1 |
                                        pff_sackAllowed == 1  |
                                        pff_hitAllowed == 1) ~ 1,
                                     (pff_hurryAllowed == 0 &
                                        pff_sackAllowed == 0  &
                                        pff_hitAllowed == 0) ~ 0)) |>
  group_by(pff_positionLinedUp) |>
  summarise(pressure_rate = mean(pressureAllowed), pressures = sum(pressureAllowed),
            sack_rate = mean(pff_sackAllowed), sacks = sum(pff_sackAllowed), plays = n()) |>
  arrange(-pressure_rate) |>
  filter(pressure_rate > 0,
         plays >= 50)
sr_pr
```

```
## # A tibble: 16 x 6
```

```
##    pff_positionLinedUp pressure_rate pressures sack_rate sacks plays
##    <chr>                       <dbl>     <dbl>     <dbl> <int> <int>
##  1 HB-L                       0.126         75   0.0202     12   595
##  2 HB-R                       0.119         68   0.0228     13   570
##  3 SRWR                       0.107          6   0           0    56
##  4 TE-R                       0.0960        46   0.0167      8   479
##  5 FB-R                       0.0909         5   0           0    55
##  6 HB                         0.085         34   0.0225      9   400
##  7 TE-L                       0.0808        27   0.0299     10   334
##  8 TE-oR                      0.08          12   0.00667     1   150
##  9 RT                         0.0771       660   0.0103     88  8555
## 10 TE-iR                      0.0731        16   0.0183      4   219
## 11 LT                         0.0726       621   0.00923    79  8556
## 12 RG                         0.0645       552   0.00584    50  8557
## 13 LG                         0.0560       479   0.00561    48  8557
## 14 TE-iL                      0.0417         5   0.025       3   120
## 15 C                          0.0414       354   0.00467    40  8557
## 16 TE-oL                      0.0152         1   0           0    66
```

```
ggplot(sr_pr, aes(sack_rate, pressure_rate)) +
  labs(
    title = 'Offensive Sack Rate vs Offensive Pressure Rate',
    subtitle = 'A look into stability | Min. 50 Pass Blocking snaps | Positive Offensive Pressure Rate'
    caption = 'By: Sam Burch',
    x = 'Offensive Sack Rate',
    y = 'Offensive Pressure Rate'
  ) +
  geom_point(aes(size = plays)) +
  geom_text_repel(aes(label = pff_positionLinedUp), size = 3) +
  stat_smooth(formula = y ~ x, method = 'lm', geom = 'line', se=FALSE, color='gray')
```



Right away, we can see how the regression line isn't a very good fit. This is due to the several reasons. One, there are small samples at different positions. Another reason, is the position tackles are put in versus interior o-lineman. Although this is the case, within the o-lineman, it is clear there is a strong positive correlation between the two metrics.

O-line Performance by Players

We will now look at the pass blocking performance across all o-lineman. Specifically, o-lineman with more than 100 pass blocking snaps and with a positive (non-zero) sack and pressure rate.

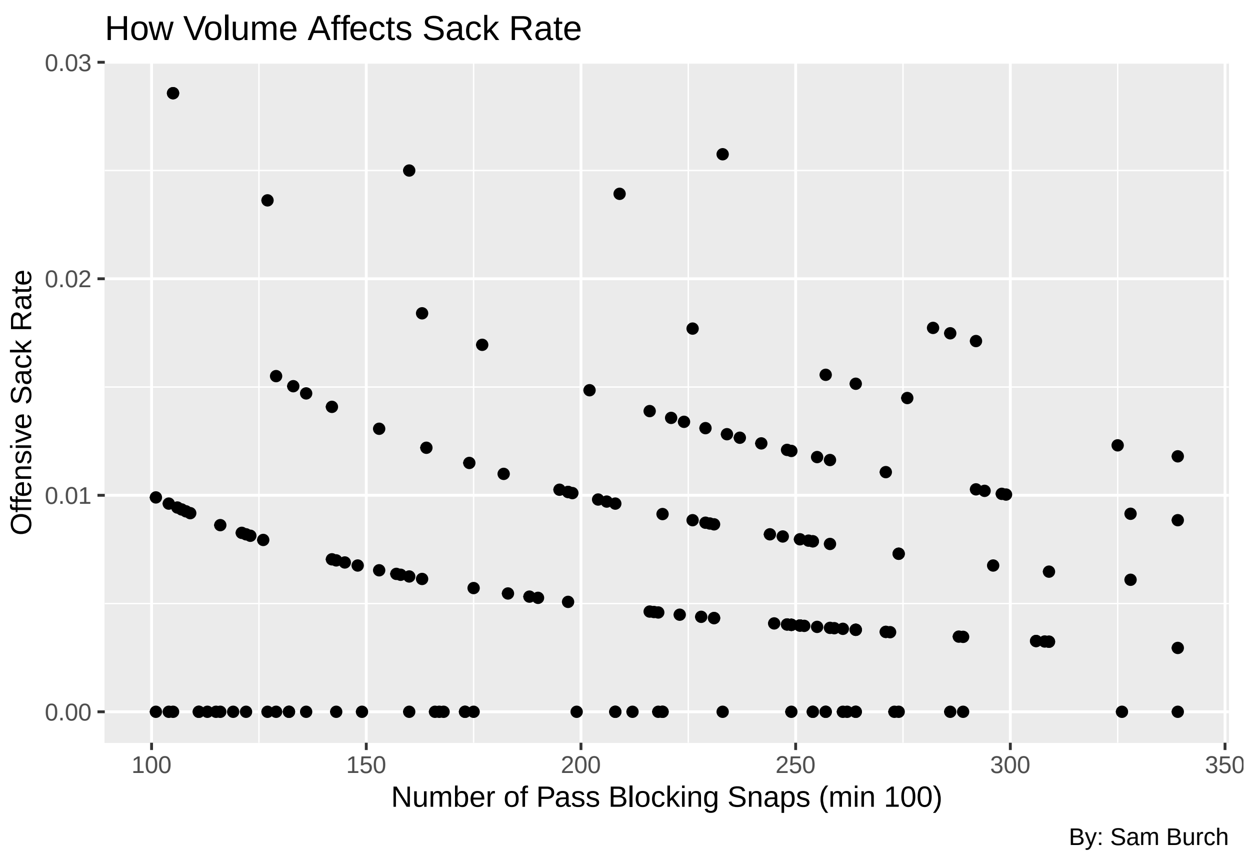
```
high_vol = pff |>
  filter(pff_role == 'Pass Block', pff_positionLinedUp == 'C'  |
                                   pff_positionLinedUp == 'LG' |
                                   pff_positionLinedUp == 'RG' |
                                   pff_positionLinedUp == 'LT' |
                                   pff_positionLinedUp == 'RT') |>
  mutate(pressureAllowed = case_when((pff_hurryAllowed == 1 |
                                        pff_sackAllowed == 1  |
                                        pff_hitAllowed == 1) ~ 1,
                                     (pff_hurryAllowed == 0 &
                                        pff_sackAllowed == 0  &
                                        pff_hitAllowed == 0) ~ 0)) |>
  group_by(nflId, pff_positionLinedUp) |>
  summarise(pressure_rate = mean(pressureAllowed), pressures = sum(pressureAllowed),
            sack_rate = mean(pff_sackAllowed), sacks = sum(pff_sackAllowed), plays = n(), .groups = 'dr
```

```r
  arrange(-pressure_rate) |>
  filter(plays >= 100)

high_vol = high_vol |>
  left_join(players, by = 'nflId') |>
  select(displayName, everything()) |>
  arrange(-pressure_rate)

high_vol
```

```
## # A tibble: 176 x 13
##    displayName   nflId pff_p~1 press~2 press~3 sack_~4 sacks plays height weight
##    <chr>         <int> <chr>     <dbl>   <dbl>   <dbl> <int> <int> <chr>   <int>
##  1 Austin Jacks~ 52426 LT        0.159      21 0           0   132 6-6       310
##  2 Cody Ford     47821 RG        0.146      18 0.00813     1   123 6-3       329
##  3 Brandon Park~ 46134 RT        0.144      15 0.00962     1   104 6-8       320
##  4 Jesse Davis   42924 RT        0.137      28 0.00980     2   204 6-6       325
##  5 Jedrick Wills 52418 LT        0.128      19 0           0   149 6-5       320
##  6 Storm Norton  45630 RT        0.126      31 0.00810     2   247 6-7       317
##  7 Liam Eichenb~ 53471 RT        0.124      13 0.0286      3   105 6-6       308
##  8 Rashod Hill   43640 LT        0.116      19 0.0122      2   164 6-6       313
##  9 Justin Herron 52603 RT        0.116      14 0.00826     1   121 6-5       290
## 10 Matt Nelson   48455 RT        0.113      37 0.00915     3   328 6-7       299
## # ... with 166 more rows, 3 more variables: birthDate <chr>, collegeName <chr>,
## #   officialPosition <chr>, and abbreviated variable names
## #   1: pff_positionLinedUp, 2: pressure_rate, 3: pressures, 4: sack_rate
```

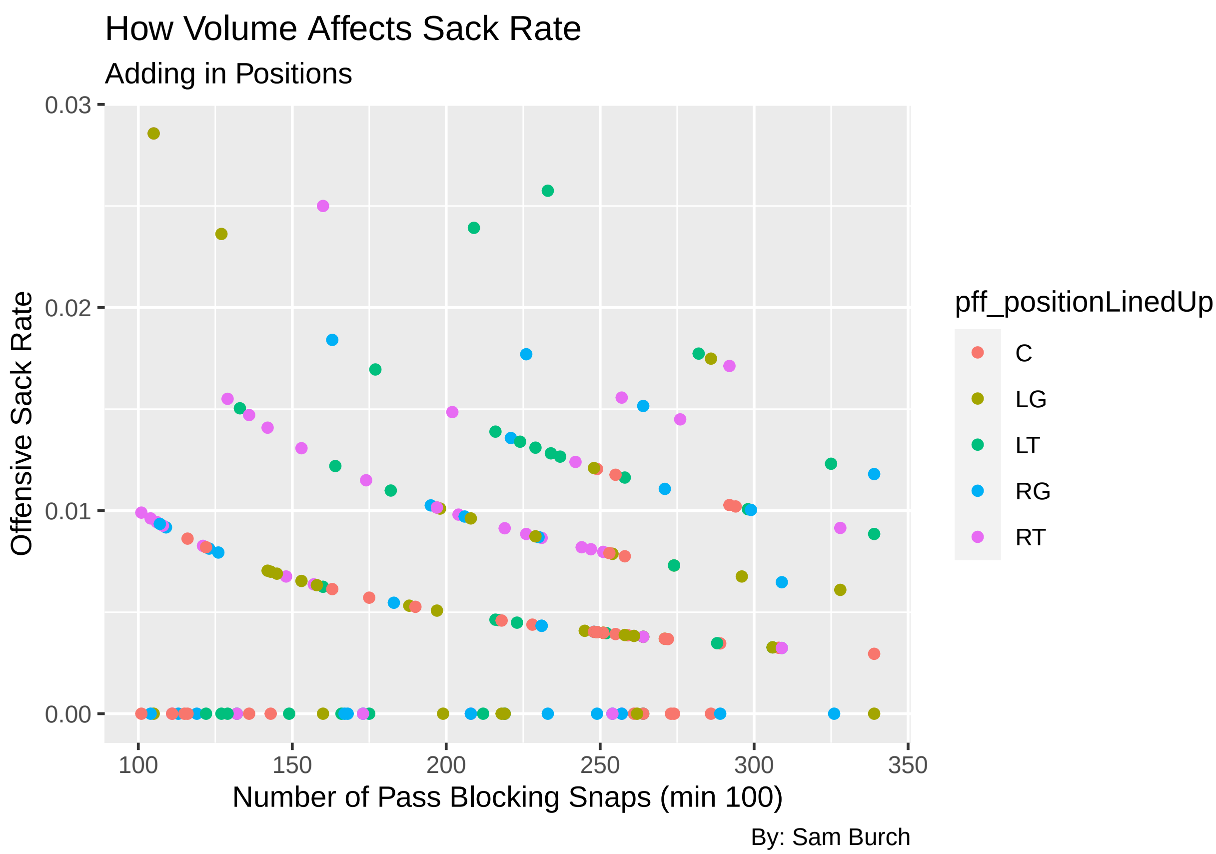
How does volume affect sack rate?

```r
ggplot(high_vol, aes(plays, sack_rate)) +
  geom_point() +
  labs(
    title = 'How Volume Affects Sack Rate',
    caption = 'By: Sam Burch',
    y = 'Offensive Sack Rate',
    x = 'Number of Pass Blocking Snaps (min 100)'
  )
```



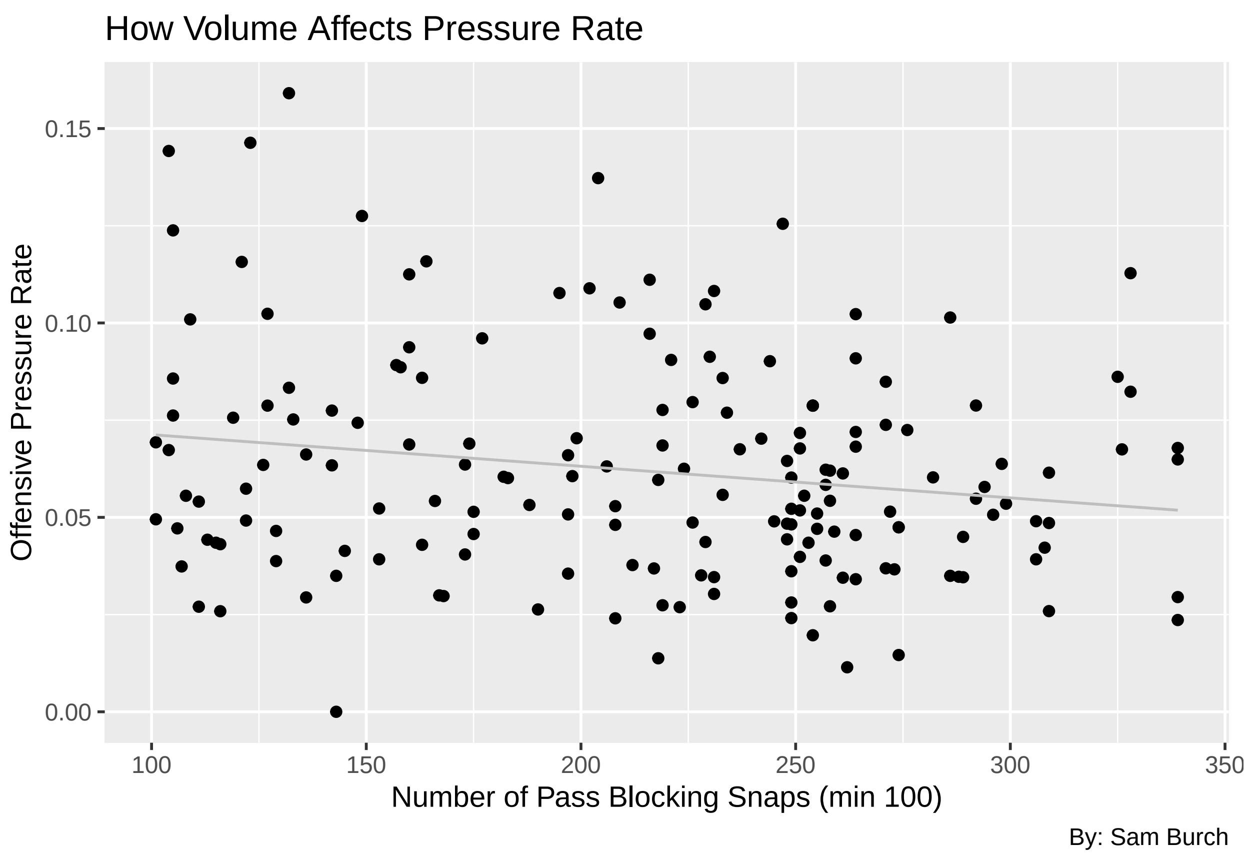
This chart does illustrate the fact that the more snaps, the lower the rate is. What is interesting here though is the different line segments. One possible scenario is this is due to the different positions; this is wrong though as one can see from the chart below.

```
ggplot(high_vol, aes(plays, sack_rate)) +
  geom_point(aes(color = pff_positionLinedUp)) +
  labs(
    title = 'How Volume Affects Sack Rate',
    subtitle = 'Adding in Positions',
    caption = 'By: Sam Burch',
    y = 'Offensive Sack Rate',
    x = 'Number of Pass Blocking Snaps (min 100)'
  )
```



The only other cause I could think of was potentially quality of play. However, this is likely wrong as good players would very likely not be this separated. So, we must note there may be some conflicting variable and move onto how volume affects offensive pressure rate.

```
ggplot(high_vol, aes(plays, pressure_rate)) +
  geom_point() +
  labs(
    title = 'How Volume Affects Pressure Rate',
    caption = 'By: Sam Burch',
    y = 'Offensive Pressure Rate',
    x = 'Number of Pass Blocking Snaps (min 100)'
  ) +
  stat_smooth(formula = y ~ x, method = 'lm', geom = 'line', se=FALSE, color='gray')
```

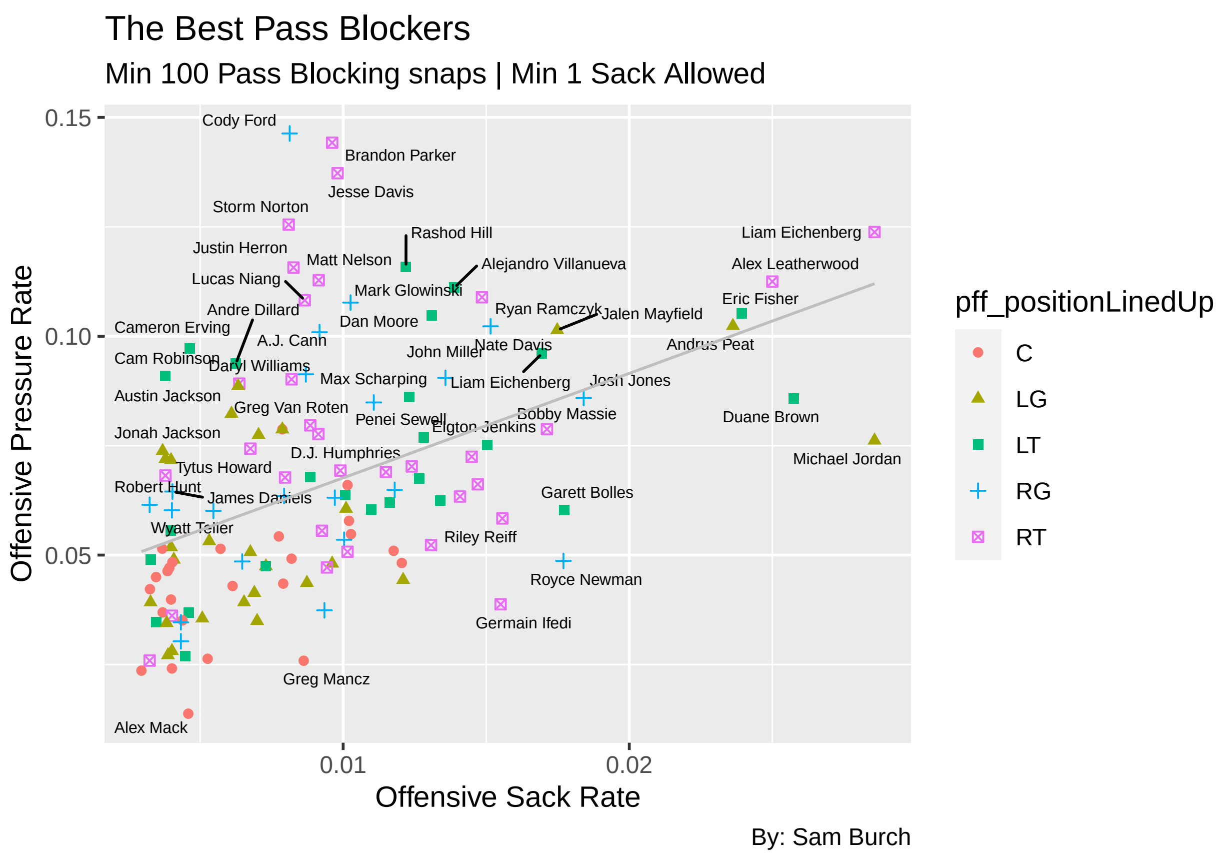


The negative correlation is still prevalent in this chart. However, the trends are not nearly as clear, and seem more random at first glance. Perhaps with pressures occurring more often than sacks, this is the reason for the more random appearance. Either way, this helps affirm that rates decrease as volume increases.

We can now look at the comparison between the rates to see the best pass-blocking players from this time frame. (Because of the number of players, several names may be cutoff.)

```
high_vol = high_vol |>
  filter(sack_rate > 0)

ggplot(high_vol, aes(sack_rate, pressure_rate)) +
  labs(
    title = 'The Best Pass Blockers',
    subtitle = 'Min 100 Pass Blocking snaps | Min 1 Sack Allowed',
    caption = 'By: Sam Burch',
    x = 'Offensive Sack Rate',
    y = 'Offensive Pressure Rate'
  ) +
  geom_point(aes(shape = pff_positionLinedUp, color = pff_positionLinedUp)) +
  geom_text_repel(aes(label = displayName), size = 2) +
  stat_smooth(formula = y ~ x, method = 'lm', geom = 'line', se=FALSE, color='gray')
```

```
## Warning: ggrepel: 83 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



This graph tells us the most useful information so far. By looking at this comparison, we can use the regression line to see who has over performed and under performed. A player above the regression line – like Alejandro Villanueva – has under performed. Since pressure rate is predictive of sack rate, the higher pressure rate relative to sack rate suggests the sack rate in the future will be higher than it is now. On the other hand, a player below the regression line – like Garrett Bolles – has over performed. Thus, he will likely regress towards a lower sack rate.

Also, o-line players towards the bottom-left have been better pass blockers than those in the upper-right. Therefore, someone like Dion Dawkins is having a much better season than Alex Leatherwood. Although this is the case, there are still a couple notes to be aware of. There is still a bias towards interior o-line, as they generally are at the bottom-left; this bias was discussed earlier. On top of that, the number of pass block reps is not displayed here. While everyone has a high amount (above 100) there is still some disparity here, so just be careful.

Defense

Total Sacks and Pressures Forced

Transitioning to the defense side, we will go through a similar process. To start off, we will look at the accumulation of sacks and pressures by defenders.

```
pff |>
  filter(pff_role == 'Pass Rush', pff_sack == 1) |>
  group_by(pff_positionLinedUp) |>
  summarise(sacks = n()) |>
  arrange(-sacks)
```

```
## # A tibble: 21 x 2
##    pff_positionLinedUp sacks
##    <chr>               <int>
##  1 LEO                   106
##  2 LOLB                   85
##  3 REO                    85
##  4 ROLB                   75
##  5 DRT                    44
##  6 DLT                    37
##  7 RE                     36
##  8 LE                     35
##  9 RILB                   19
## 10 LLB                    15
## # ... with 11 more rows
```

```
pff |>
  filter(pff_role == 'Pass Rush',
         pff_hit == 1 |
         pff_hurry == 1 |
         pff_sack == 1) |>
  group_by(pff_positionLinedUp) |>
  summarise(pressures = n()) |>
  arrange(-pressures)
```

```
## # A tibble: 26 x 2
##    pff_positionLinedUp pressures
##    <chr>                   <int>
##  1 LEO                       623
##  2 LOLB                      580
##  3 REO                       521
##  4 ROLB                      479
##  5 DRT                       381
##  6 LE                        357
##  7 DLT                       354
##  8 RE                        259
##  9 NT                        122
## 10 LILB                       94
## # ... with 16 more rows
```

There are a lot more positions being considered on the defensive side – since everyone can rush the passer. With that being said, we see a similarity to the offensive side of the ball; lineman are higher than other positions in these metrics. We can even see how edge rushers have higher numbers than interior d-lineman. These make sense, because of similarity, so we can move on.

Defensive Sack and Pressure Rates

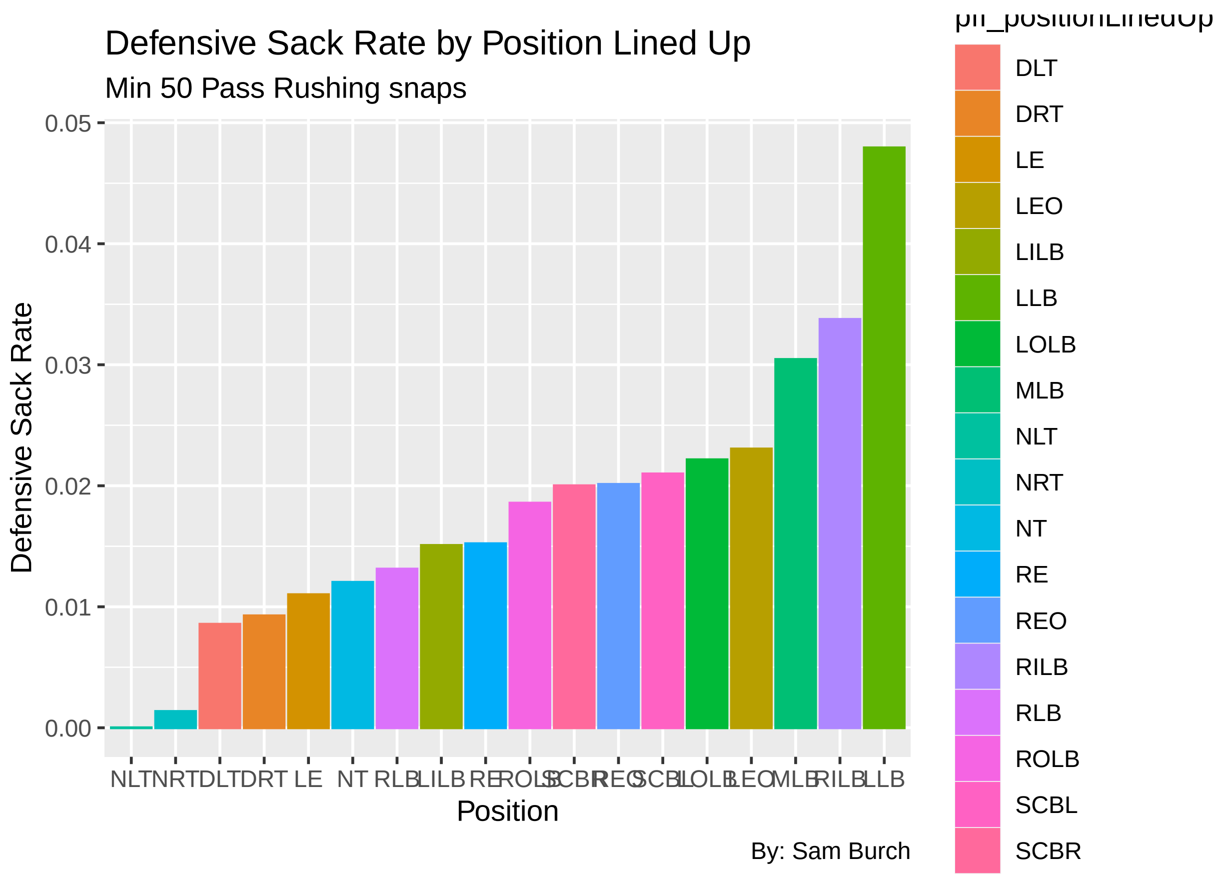
Let us add in defensive sack and pressure rates.

```
def_rates = pff |>
  filter(pff_role == 'Pass Rush') |>
  mutate(pressure = case_when((pff_hurry == 1 |
                               pff_sack == 1  |
```

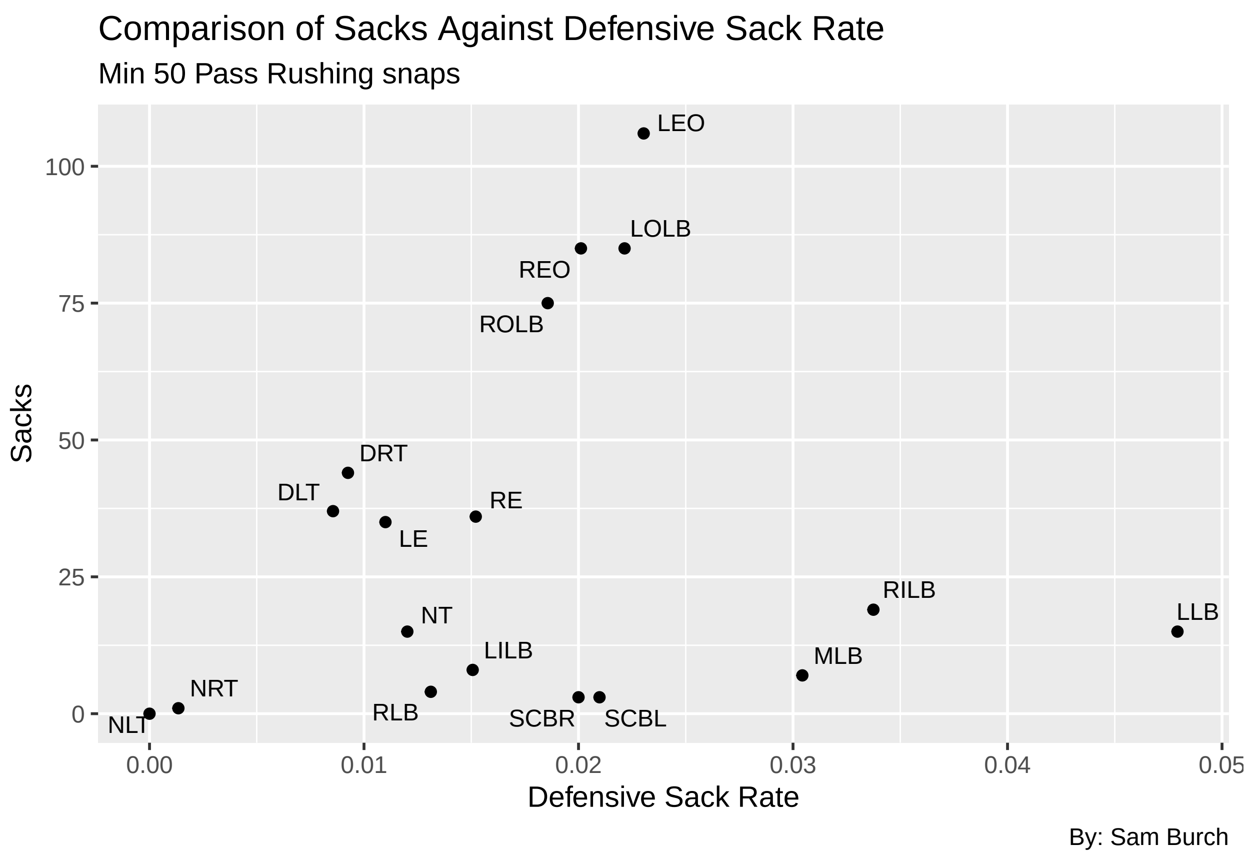
```r
                              pff_hit == 1) ~ 1,
                             (pff_hurry == 0 &
                              pff_sack == 0  &
                              pff_hit == 0) ~ 0)) |>
  group_by(pff_positionLinedUp) |>
  summarise(pressure_rate = mean(pressure), pressures = sum(pressure),
            sack_rate = mean(pff_sack), sacks = sum(pff_sack), plays = n()) |>
  arrange(-pressure_rate) |>
  filter(plays >= 50)
def_rates
```

```
## # A tibble: 18 x 6
##    pff_positionLinedUp pressure_rate pressures sack_rate sacks plays
##    <chr>                       <dbl>     <dbl>     <dbl> <int> <int>
##  1 MLB                        0.239         55   0.0304      7   230
##  2 SCBL                       0.224         32   0.0210      3   143
##  3 LLB                        0.211         66   0.0479     15   313
##  4 RLB                        0.193         59   0.0131      4   305
##  5 LILB                       0.177         94   0.0151      8   531
##  6 RILB                       0.163         92   0.0337     19   563
##  7 SCBR                       0.16          24   0.02        3   150
##  8 LOLB                       0.151        580   0.0221     85  3838
##  9 LEO                        0.135        623   0.0230    106  4601
## 10 REO                        0.123        521   0.0201     85  4226
## 11 ROLB                       0.119        479   0.0186     75  4040
## 12 LE                         0.112        357   0.0110     35  3182
## 13 RE                         0.109        259   0.0152     36  2367
## 14 NT                         0.0978       122   0.0120     15  1248
## 15 DLT                        0.0818       354   0.00855    37  4325
## 16 DRT                        0.0801       381   0.00925    44  4756
## 17 NLT                        0.0730        43   0           0   589
## 18 NRT                        0.0565        42   0.00135     1   743
```

```r
ggplot(def_rates, aes(reorder(pff_positionLinedUp, sack_rate), sack_rate)) +
  geom_col(aes(color = pff_positionLinedUp, fill = pff_positionLinedUp)) +
  # theme(axis.text.x = element_blank()) +
  labs(
    title = 'Defensive Sack Rate by Position Lined Up',
    subtitle = 'Min 50 Pass Rushing snaps',
    caption = 'By: Sam Burch',
    y = 'Defensive Sack Rate',
    x = 'Position'
  )
```




```
ggplot(def_rates, aes(sack_rate, sacks)) +
  geom_point() +
  geom_text_repel(aes(label = pff_positionLinedUp), size = 3) +
  labs(
    title = 'Comparison of Sacks Against Defensive Sack Rate',
    subtitle = 'Min 50 Pass Rushing snaps',
    caption = 'By: Sam Burch',
    y = 'Sacks',
    x = 'Defensive Sack Rate'
  )
```



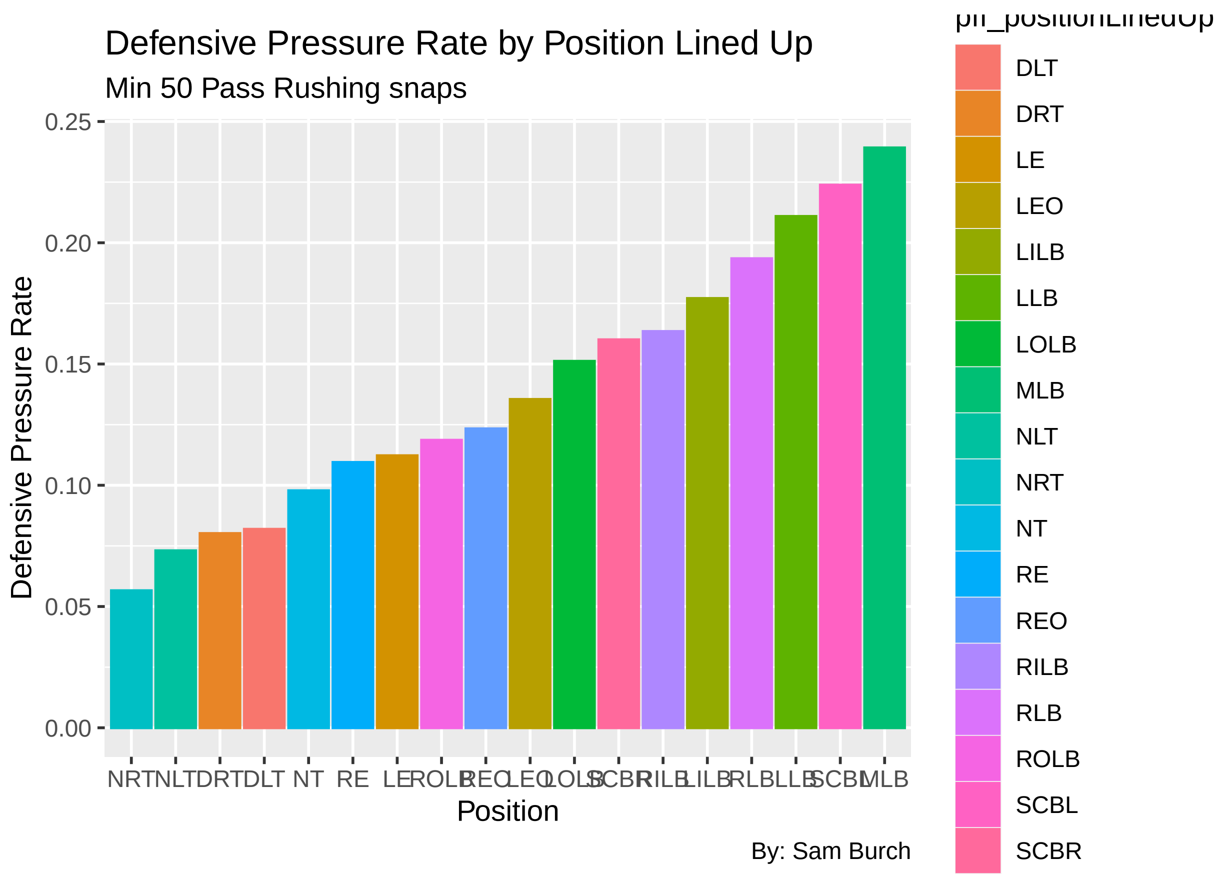
We start to see the trend of as volume increases, the rates decrease; the notable exception are nose tackles. Nose tackles jobs are rarely to straight up rush the passer. So, it makes sense that they have low sack rates, even though they have a relatively high amount of "pass rushing" snaps.

Another takeaway is the high sack rates for linebackers. While (once again) we must be aware of the smaller sample size, linebackers show up well here. A possible reason is when linebackers rush the QB, it might be a blitz more often than not. So, if d-lineman eat up most of the blocks, this gives linebackers the chance to capitalize and get to the QB.

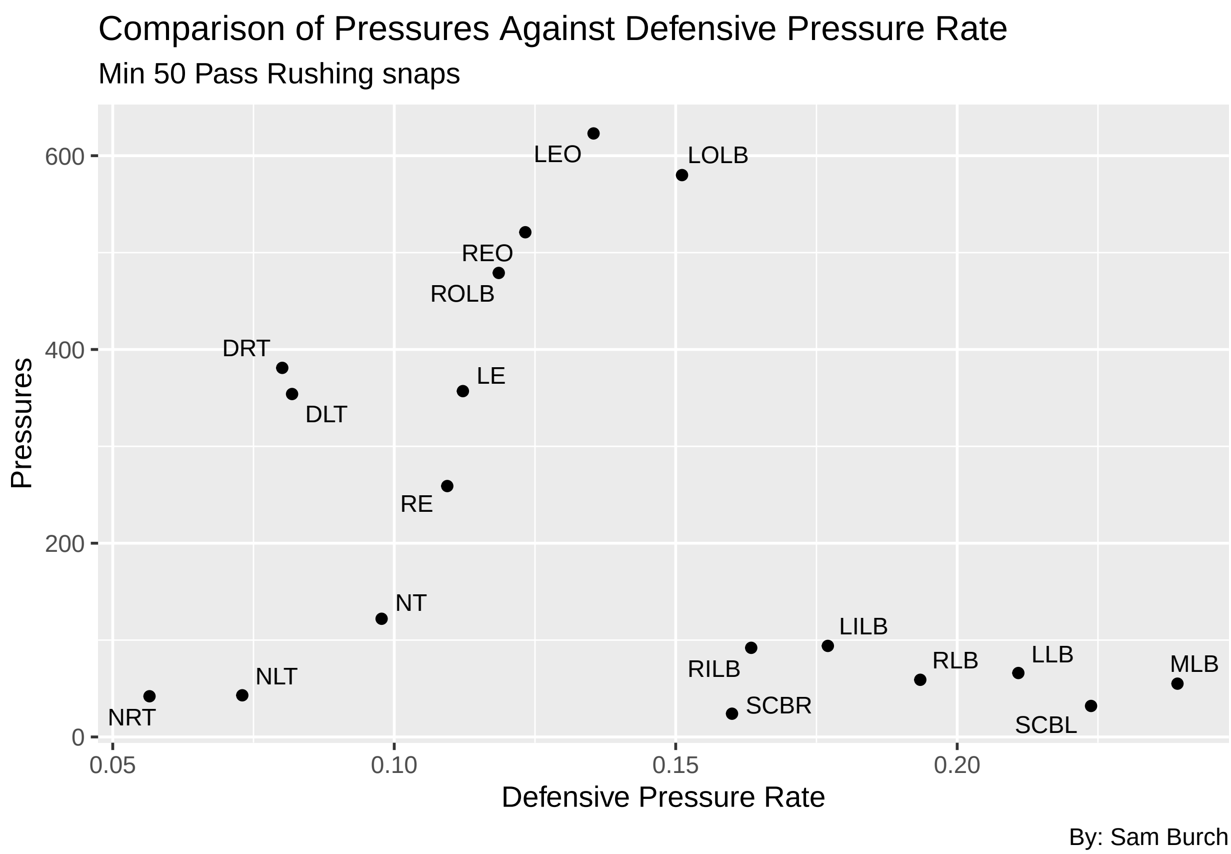
The second graph above gives similar takeaways to earlier. Note that players on the left of the d-line have slightly better success, which again may be due to the blindside effect.

Onto illustrating defensive pressure rate.

```r
ggplot(def_rates, aes(reorder(pff_positionLinedUp, pressure_rate), pressure_rate)) +
  geom_col(aes(color = pff_positionLinedUp, fill = pff_positionLinedUp)) +
  # theme(axis.text.x = element_blank()) +
  labs(
    title = 'Defensive Pressure Rate by Position Lined Up',
    subtitle = 'Min 50 Pass Rushing snaps',
    caption = 'By: Sam Burch',
    y = 'Defensive Pressure Rate',
    x = 'Position'
  )
```




```r
ggplot(def_rates, aes(pressure_rate, pressures)) +
  geom_point() +
  geom_text_repel(aes(label = pff_positionLinedUp), size = 3) +
  labs(
    title = 'Comparison of Pressures Against Defensive Pressure Rate',
    subtitle = 'Min 50 Pass Rushing snaps',
    caption = 'By: Sam Burch',
    y = 'Pressures',
    x = 'Defensive Pressure Rate'
  )
```



Because of the similarity to defensive sack rate, the analysis above is suffice here as well.

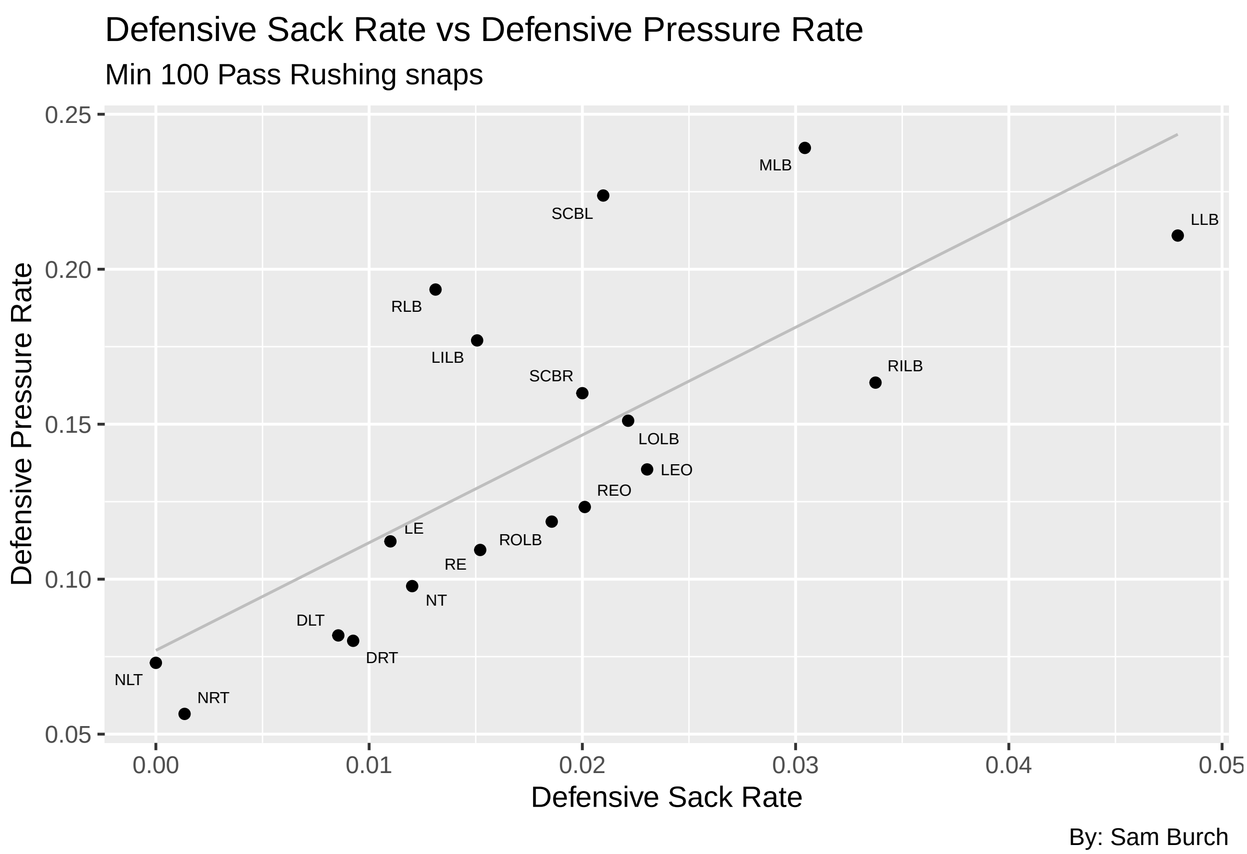
Defensive Sack Rate vs Defensive Pressure Rate

```
def_high = def_rates |>
  filter(plays >= 100)
def_high
```

```
## # A tibble: 18 x 6
##    pff_positionLinedUp pressure_rate pressures sack_rate sacks plays
##    <chr>                       <dbl>     <dbl>     <dbl> <int> <int>
##  1 MLB                         0.239        55    0.0304     7   230
##  2 SCBL                        0.224        32    0.0210     3   143
##  3 LLB                         0.211        66    0.0479    15   313
##  4 RLB                         0.193        59    0.0131     4   305
##  5 LILB                        0.177        94    0.0151     8   531
##  6 RILB                        0.163        92    0.0337    19   563
##  7 SCBR                        0.16         24    0.02       3   150
##  8 LOLB                        0.151       580    0.0221    85  3838
##  9 LEO                         0.135       623    0.0230   106  4601
## 10 REO                         0.123       521    0.0201    85  4226
## 11 ROLB                        0.119       479    0.0186    75  4040
## 12 LE                          0.112       357    0.0110    35  3182
## 13 RE                          0.109       259    0.0152    36  2367
```

```
## 14 NT                              0.0978        122   0.0120      15   1248
## 15 DLT                             0.0818        354   0.00855     37   4325
## 16 DRT                             0.0801        381   0.00925     44   4756
## 17 NLT                             0.0730         43   0            0    589
## 18 NRT                             0.0565         42   0.00135      1    743
```

```r
ggplot(def_high, aes(sack_rate, pressure_rate)) +
  geom_point() +
  geom_text_repel(aes(label = pff_positionLinedUp), size = 2) +
  labs(
    title = 'Defensive Sack Rate vs Defensive Pressure Rate',
    subtitle = 'Min 100 Pass Rushing snaps',
    caption = 'By: Sam Burch',
    x = 'Defensive Sack Rate',
    y = 'Defensive Pressure Rate'
  ) +
  stat_smooth(formula = y ~ x, method = 'lm', geom = 'line', se=FALSE, color='gray')
```



This graphs illustrates that d-lineman have been lucky, but that's not the case. Again, there are biases with small sample sizes among the other positions that affect the regression line. Just looking at d-lineman, the strong positive correlation makes sense. Thus, we will focus on these positions going forwards, as the other positions just add in too much noise.

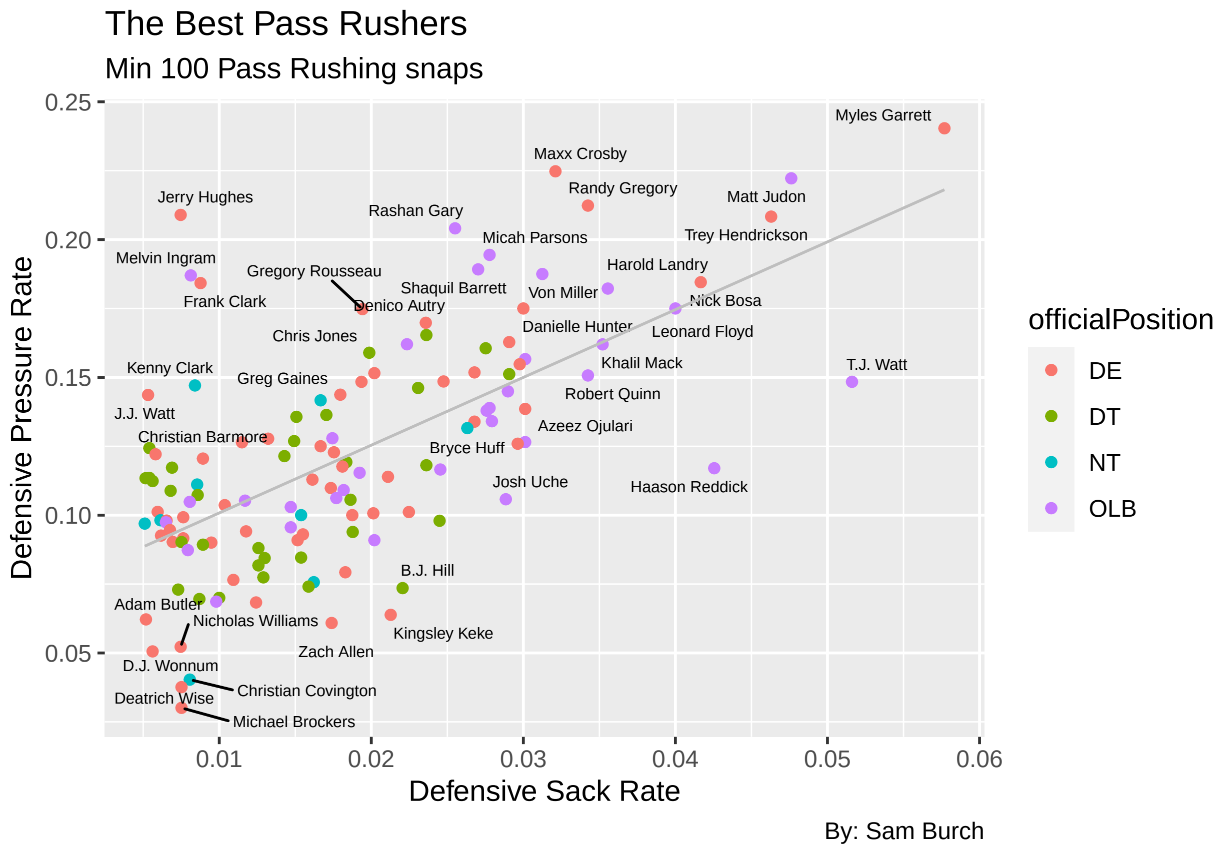
D-line Performance by Players

```
def_vol = pff |>
  filter(pff_role == 'Pass Rush') |>
  mutate(pressure = case_when((pff_hurry == 1 |
                                pff_sack == 1  |
                                pff_hit == 1) ~ 1,
                             (pff_hurry == 0 &
                                pff_sack == 0  &
                                pff_hit == 0) ~ 0)) |>
  group_by(nflId) |>
  summarise(pressure_rate = mean(pressure), pressures = sum(pressure),
            sack_rate = mean(pff_sack), sacks = sum(pff_sack), plays = n()) |>
  arrange(-pressure_rate) |>
  filter(plays >= 100, sack_rate > 0) |>
  left_join(players, by = 'nflId') |>
  select(displayName, everything())
def_vol
```

```
## # A tibble: 132 x 12
##    displayName   nflId press~1 press~2 sack_~3 sacks plays height weight birth~4
##    <chr>         <int>   <dbl>   <dbl>   <dbl> <int> <int> <chr>   <int> <chr>
##  1 Myles Garrett 44813   0.240      50 0.0577     12   208 6-4       272 1995-1~
##  2 Maxx Crosby   47889   0.225      49 0.0321      7   218 6-5       255 1997-0~
##  3 Matt Judon    43435   0.222      42 0.0476      9   189 6-3       261 1992-0~
##  4 Randy Gregory 42403   0.212      31 0.0342      5   146 6-5       255 1992-1~
##  5 Jerry Hughes  35470   0.209      28 0.00746     1   134 6-2       254 1988-0~
##  6 Trey Hendric~ 44915   0.208      45 0.0463     10   216 6-4       270 1994-1~
##  7 Rashan Gary   47795   0.204      40 0.0255      5   196 6-5       277 1997-1~
##  8 Micah Parsons 53441   0.194      21 0.0278      3   108 6-3       245 <NA>
##  9 Shaquil Barr~ 41915   0.189      42 0.0270      6   222 6-2       250 1992-1~
## 10 Von Miller    37075   0.188      30 0.0312      5   160 6-3       250 1989-0~
## # ... with 122 more rows, 2 more variables: collegeName <chr>,
## #   officialPosition <chr>, and abbreviated variable names 1: pressure_rate,
## #   2: pressures, 3: sack_rate, 4: birthDate
```

```
ggplot(def_vol, aes(sack_rate, pressure_rate)) +
  geom_point(aes(color = officialPosition)) +
  geom_text_repel(aes(label = displayName), size = 2) +
  stat_smooth(formula = y ~ x, method = 'lm', geom = 'line', se=FALSE, color='gray') +
  labs(
    title = 'The Best Pass Rushers',
    subtitle = 'Min 100 Pass Rushing snaps',
    caption = 'By: Sam Burch',
    x = 'Defensive Sack Rate',
    y = 'Defensive Pressure Rate'
  )
```

```
## Warning: ggrepel: 93 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



We are left with our final chart. This discusses the best pass rushers. As I explained how one can read the best pass blockers chart, this is very similar. The only difference is it is flipped in terms of quality players. This is because higher sack rates and pressure rates are good for the defensive players.

Hence, players like Melvin Ingram and Rashan Gary have gotten unlucky. On the other hand, players like Hasson Reddick and T.J. Watt have gotten lucky. Meanwhile, Myles Garrett has been utterly dominant and Jerry Tillery not so much.

Lastly, we will take note of the interior d-lineman showing up worse than the edge defenders. This, again, is because edge defenders are better set up to rush the passer.

Conclusion

By looking at the data, we found a way to measure how good pass blockers are and how good pass rushers are. Combining pressure rate and sack rate doesn't give the perfect answer to who these best players are, but it gives us a good idea. If someone just looks at the top sacks by position, there can be so much noise where it won't tell you much. The goal is to find a good approximation with little noise. Then, one can adjust based off of other factors, if needed. We have accomplished that here because these rates help eliminate noise and show a good approximation of the impact these players have had on improving their team's chances of winning.

I would like to thank you for your consideration. Analytics have helped turn complex problems in football into seemingly simple answers. Because of this, football workers, players, and fans has gotten continually smarter over the years. I hope this helps contribute to that in one way or another!

References

- [The Washington Post] (https://www.washingtonpost.com/news/fancy-stats/wp/2017/08/01/the-value-of-a-sack-and-why-pass-rusher-is-the-nfls-second-most-important-position/)
- [The Big Lead] (https://www.thebiglead.com/posts/sacks-are-a-quarterback-stat-01dxqapkgvw9)
- [PFF] (https://www.pff.com/news/pro-the-importance-of-pressure-its-not-all-about-sacks)