

544_Final_Project_NFL

Sufyan Hammoudeh

11/14/2023

Class: SI 544 Name: Sufyan Hammoudeh

Research Statement: I am a graduate student who loves to watch and play sports. That is why I have decided for my Final Project to study yearly offensive NFL Stats from 2012-2022. Also, as a Lions fan, I am looking to find any players stats for the Detroit Lions. I hope to gain some more insight on the raw data provided.

```
library(tidyverse)

## — Attaching packages — tidyverse
1.3.2 —✓ ggplot2 3.3.6    ✓ purrr 0.3.4
## ✓ tibble 3.1.7    ✓ dplyr 1.0.9
## ✓ tidyr 1.2.0    ✓ stringr 1.4.0
## ✓ readr 2.1.2    ✓ forcats 0.5.1 — Conflicts
                                tidyverse_conflicts() —
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()

library(dplyr)
library(infer)
library(moderndiver)
library(ggplot2)


# First I will pull all the offensive NFL yearly stats from 2012-2022
nfl <- read.csv("yearly_data_updated_08_23.csv")

# The following code shows the number of players recorded in each position
# per season
pos_groups <- nfl %>%
  group_by(season, position) %>%
  count(id) %>%
  summarise(count = n())

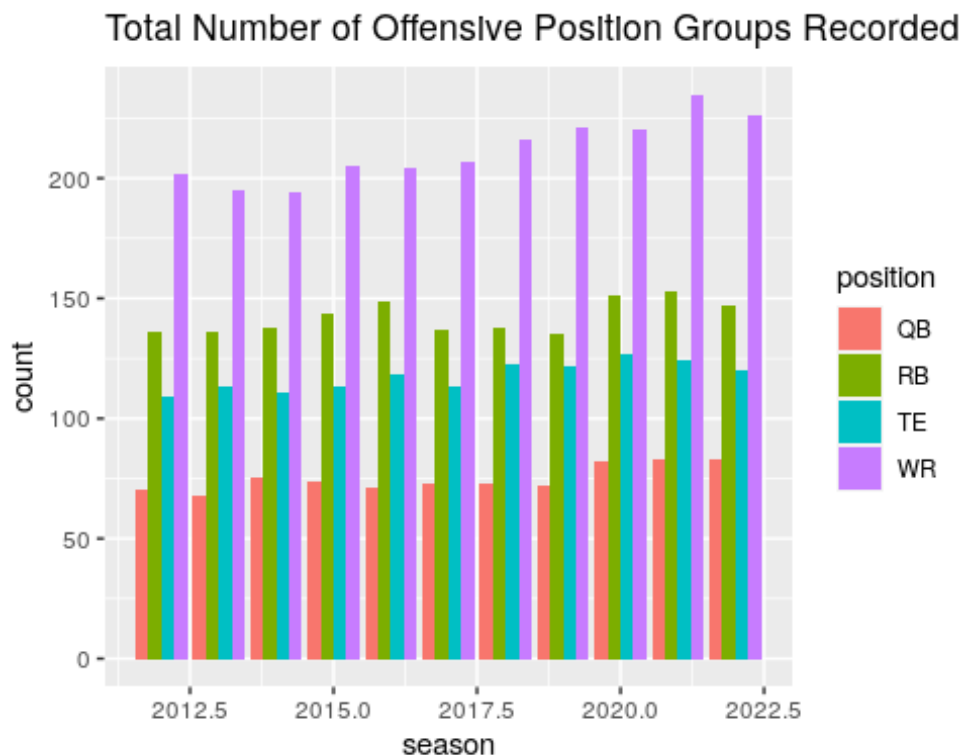
## `summarise()` has grouped output by 'season'. You can override using the
## `.groups` argument.

pos_groups

## # A tibble: 44 × 3
## # Groups:   season [11]
##   season position count
##   <int> <chr>    <int>
## 1    2012 QB         70
```

```
## 2    2012 RB      136
## 3    2012 TE      109
## 4    2012 WR      202
## 5    2013 QB       68
## 6    2013 RB      136
## 7    2013 TE      113
## 8    2013 WR      195
## 9    2014 QB       75
## 10   2014 RB      138
## # ... with 34 more rows
## #  Use `print(n = ...)` to see more rows

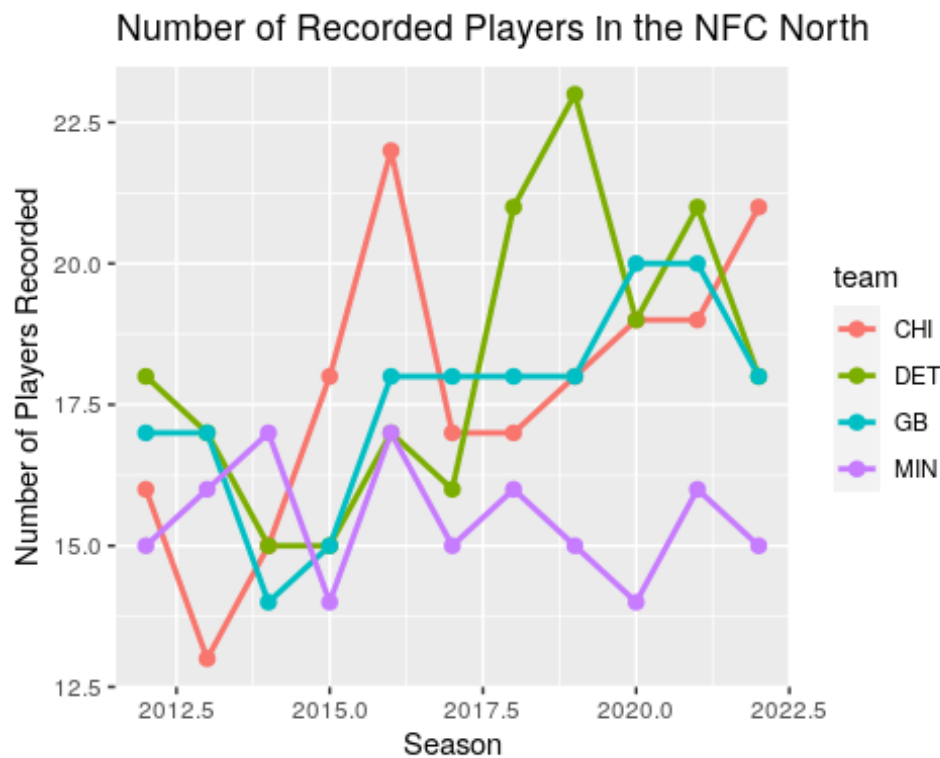
#visulazation
pos_groups %>%
  ggplot(aes(x= season, y= count, fill = position)) +
  geom_col(position = "dodge")+
  labs(title="Total Number of Offensive Position Groups Recorded per Season")
```



The data above shows the number of NFL offensive positions reported per each season. Fun fact, all of these positions are called skilled position in all levels of football. The labels for each one stand for Quarterback (QB), Running back (RB), Tight End (TE), and Wide Receiver (WR). As we can see, the wide receivers make up for most of the recorded data due to so many of them player. Typically, there are about three WR, one to two RBs and TEs, and one QB on the field at all times. There are also more substitutions for WRs than any other group. Based on the information we have, we can see WR recorded stats have steadily increased per each season compared to other positions. This could mean one of

two things. One, is injuries, which are very common among all of these positions but more so with WRs. Two, depending what kind of offensive style each team has, they could be more of a run team or a pass. Many teams have moved more to a passing scheme than running scheme which includes more WRs along with TEs and RBs.

```
pos_nfc_north <- nfl %>%  
  filter(team %in% c("DET", "GB", "MIN", "CHI")) %>%  
  group_by(season, team) %>%  
  summarise(participants = n())  
  
## `summarise()` has grouped output by 'season'. You can override using the  
## `.groups` argument.  
  
pos_nfc_north %>%  
  ggplot(aes(x = season, y = participants, group=team))+  
  geom_line(aes(color=team), size =1)+  
  geom_point(aes(color=team), size =2.3)+  
  labs(x = "Season", y = "Number of Players Recorded", title = "Number of  
Recorded Players in the NFC North")
```



The NFL has two main divisions, the AFC and NFC. Within each division they have sub-divisions North, East, South, and West. I chose to do the NFC North because the Detroit Lions are in that division. Based on the graph, the Chicago Bears (CHI) have recorded the most amount of total offensive players after the 2022 season. This could be based on salary cap space to retain or acquire new players that are starters. This could also be based on how long each offensive player has played on that team, meaning they could be the only player in a certain position with recorded stats. An example of this is QB, Aaron Rodgers, who has played for

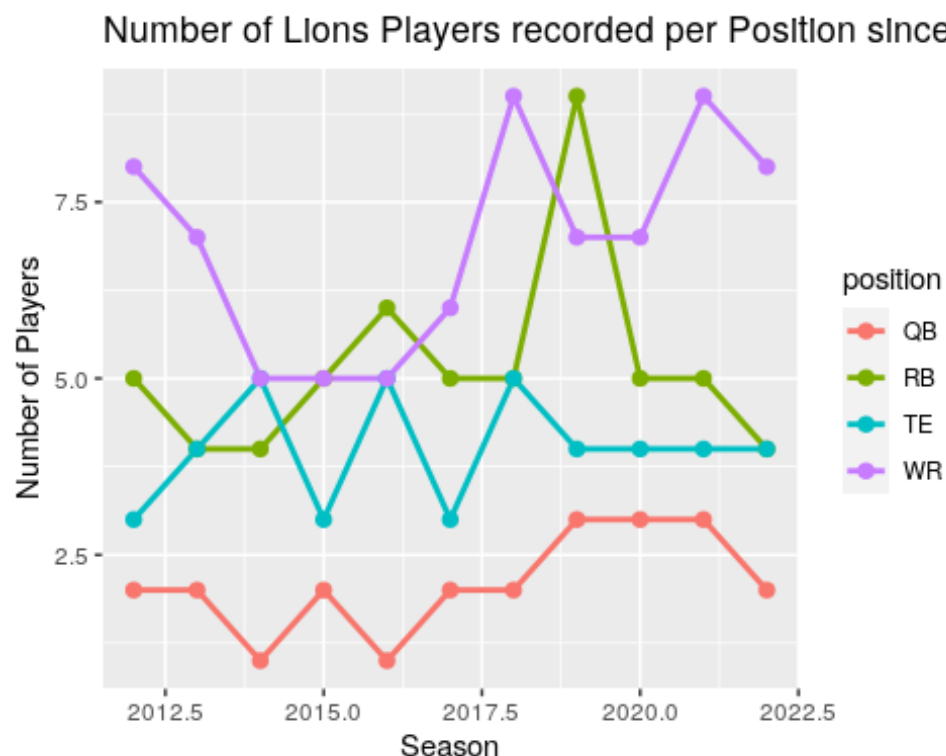
the Green Bay Packers for most of his career until he has recently signed with the New York Jets as of this year.

Now, Let's Look into the Detroit Lions (my favorite football team) and see how many offensive players they have recorded per position for every season since 2012.

```
lions_pos <- nfl %>%  
  group_by(season, team, position) %>%  
  filter(team=="DET") %>%  
  count(id) %>%  
  summarise(participants = n())
```

```
## `summarise()` has grouped output by 'season', 'team'. You can override  
using the  
## `.groups` argument.
```


```
lions_pos %>%  
  ggplot(aes(x = season, y = participants, group=position)) +  
  geom_line(aes(color=position), size = 1) +  
  geom_point(aes(color=position), size = 2.3) +  
  labs(x = "Season", y = "Number of Players", title = "Number of Lions  
Players recorded per Position since 2012 season")
```



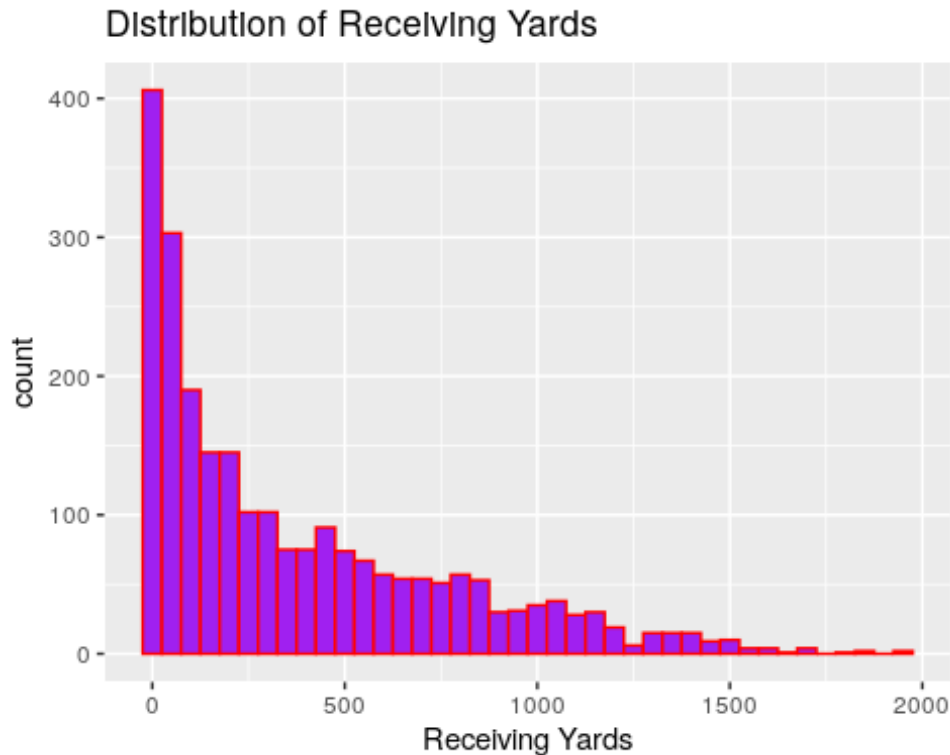
One trend that caught my attention the most was how many players stats were recorded among RBs compared to WRs during the 2018 season. I was expecting WRs to have more stats recorded than RBs during that year because the Lions had an amazing QB named Matthew Stafford. But, this was also the year where the Lions acquired a new head coach named

Matt Patricia. His style of play was finding a balance in the passing and running game, but he leaned more into the running the ball more than passing it. Based on this information, I want to see the distribution of receiving yards for all teams since the 2012-2022 season compared to the lions receiving yards between 2012-2022 season.

```
rel_receiving_yards <- nfl %>%group_by(receiving_yards)
%>%filter(team=="DET")%>%summarize(player = n())
rel_receiving_yards

## # A tibble: 143 × 2
##   receiving_yards player
##           <int> <int>
## 1             -6      1
## 2             -2      2
## 3              0     33
## 4              1      1
## 5              3      2
## 6              5      2
## 7              6      2
## 8              7      1
## 9              8      1
## 10             9      2
## # ... with 133 more rows
## #  Use `print(n = ...)` to see more rows

wr_data <- subset(nfl, position == "WR")
ggplot(wr_data,aes(x = receiving_yards)) +
  geom_histogram(binwidth = 50, color = "red", fill = "purple") +
  labs(x = "Receiving Yards", title = "Distribution of Receiving Yards")
```

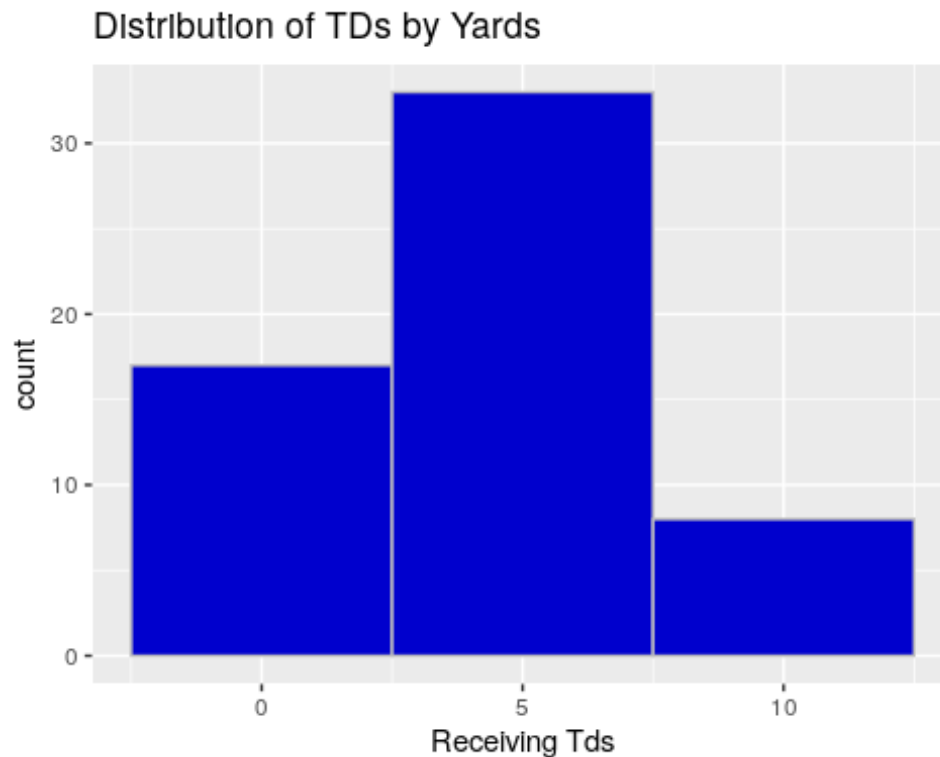


To my surprise, over 400 WRs who were targeted at least once, had between -3 and 6 receiving yards per season between 2012-2022, but not many WRs ended with negative yards. The histogram is skewed to the right, meaning it has a positive outlook. This could be due to season ending injuries, being cut from their team, or their QB just never threw the ball to them, which is unlikely but can happen. It is rare to see players go for over 750 yards per season or even 1000 yards, but there are so few super star WRs that have achieved such incredible stats. For example, star WR for the Miami Dolphins, Tyreek Hill (also known as the cheetah) is on pace to break an NFL record of having over 2000 receiving yards in the 2023 season. Check this link to see why: <https://www.nbcsports.com/nfl/profootballtalk/rumor-mill/news/tyreek-hill-remains-ahead-of-pace-to-break-nfl-record-for-receiving-yards-in-a-season>. Seeing these stats made me curious to the receiving yards distribution for the Lions who have scored one or more TD throughout the regular seasons.

```
yards_tds <- nfl %>%
  filter(team == "DET", receiving_tds >= 2) %>%
  group_by(receiving_yards, receiving_tds, team) %>%
  summarize(player = n())

## `summarise()` has grouped output by 'receiving_yards', 'receiving_tds'.
## You can
## override using the `.groups` argument.

ggplot(yards_tds, aes(x=receiving_tds)) +
  geom_histogram(binwidth = 5, color = "grey", fill = "Mediumblue") +
  labs(x = "Receiving Tds", title = "Distribution of TDs by Yards")
```



I am seeing here that there really isn't much of a difference between the receiving yards and the average number of receiving TDs for the Lions, which was between 2 and 7 TDs for each WR. But I wanted to see if the average receiving TDs for Lions WRs for every season. I conducted the code below.

```
yards_tds_1 <- nfl %>%
  filter(team == "DET") %>%
  group_by(receiving_yards, receiving_tds, team, season) %>%
  summarize(player = n())

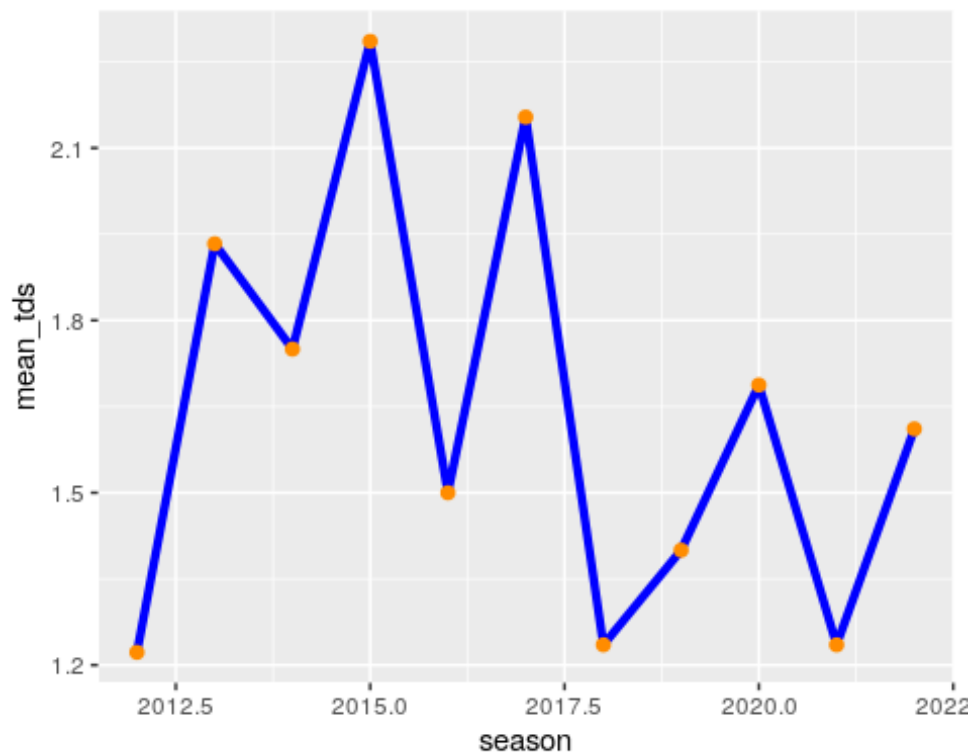
## `summarise()` has grouped output by 'receiving_yards', 'receiving_tds',
## 'team'.
## You can override using the `.groups` argument.

tds_mean <-
yards_tds_1 %>% group_by(season) %>% summarize(mean_tds = mean(receiving_tds))
tds_mean

## # A tibble: 11 × 2
##   season mean_tds
##   <int>   <dbl>
## 1  2012     1.22
## 2  2013     1.93
## 3  2014     1.75
## 4  2015     2.29
## 5  2016     1.5
## 6  2017     2.15
## 7  2018     1.24
```

```
## 8    2019    1.4
## 9    2020    1.69
## 10   2021    1.24
## 11   2022    1.61
```

```
ggplot(tds_mean,aes(x=season,y=mean_tds))+
  geom_line(size=1.5,color="blue")+
  geom_point(color="dark orange",size =2)
```

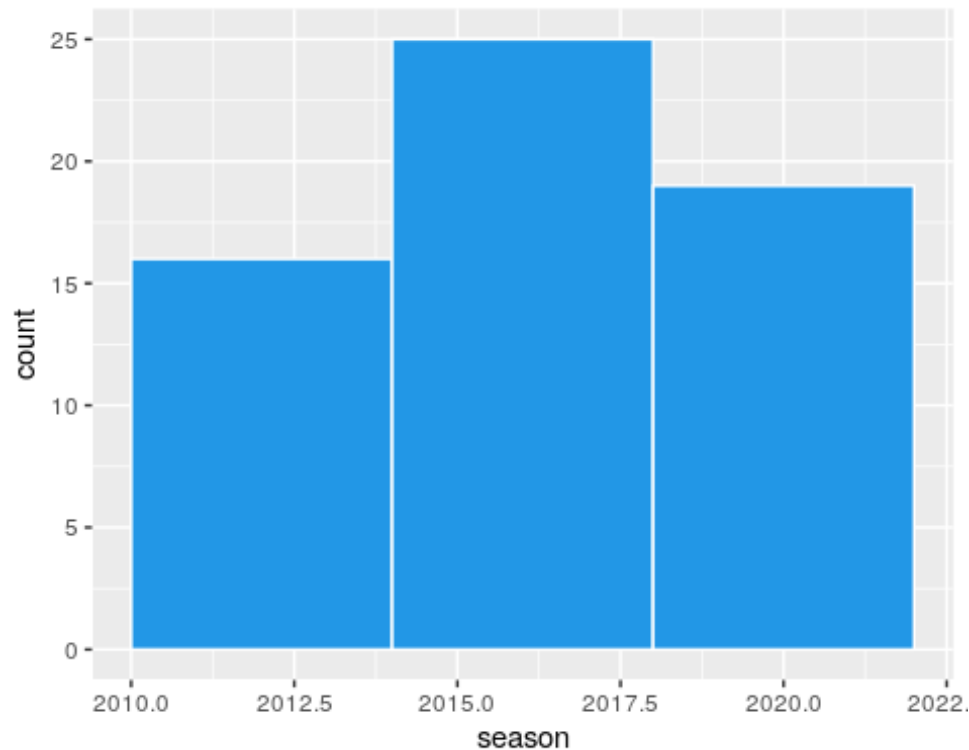


I found that to my surprise, each WR for the Lions scored around 2 TDs per season, which is on the lower end of the average from the previous graph. I wanted to see what season did the group of Lions WRs scored the most TDs.

```
player_wr_tds <- nfl %>%group_by(receiving_tds,team,season)
%>%filter(team=="DET")%>%summarize(numofage = n())

## `summarise()` has grouped output by 'receiving_tds', 'team'. You can
## override
## using the `.groups` argument.

ggplot(player_wr_tds,aes(x=season))+
  geom_histogram(binwidth = 4,color="white",fill = "2324")
```

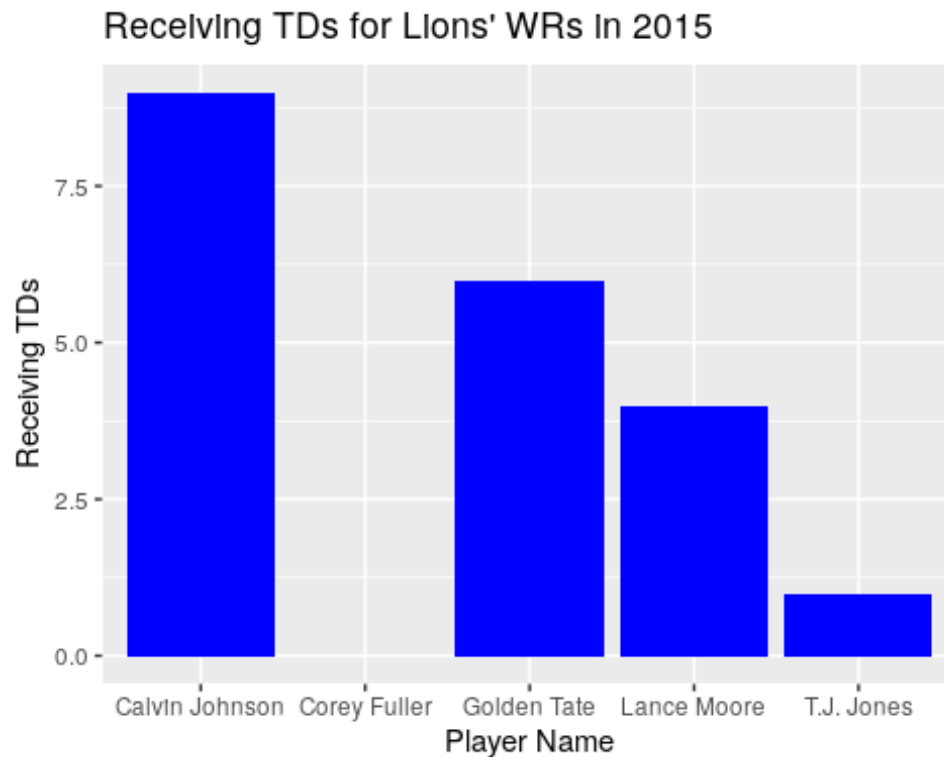



Well, would you look at that. The Lions WR group had the most TDs in around 2015 which was close to 25 TDs. Star WR and now Hall of Famer, Calvin Johnson (aka Megatron) was probably the one person who lead all WRs in TDs. Sadly, this would be Megatron's final season in the NFL as he decided to retire early at the age of 31. I wanted to test this theory to see if Mr. Megatron did lead all WRs in TDs in 2015. So, I conducted the graph below.

```
lions_wr <- nfl%>%group_by(receiving_tds, name, position,team,season)
%>%filter(team=="DET",position=="WR",season=="2015")%>%summarize(numoftds =
n())

## `summarize()` has grouped output by 'receiving_tds', 'name', 'position',
'team'.
## You can override using the `.groups` argument.

ggplot(lions_wr, aes(x = name, y = receiving_tds)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Player Name", y = "Receiving TDs", title = "Receiving TDs for
Lions' WRs in 2015")
```



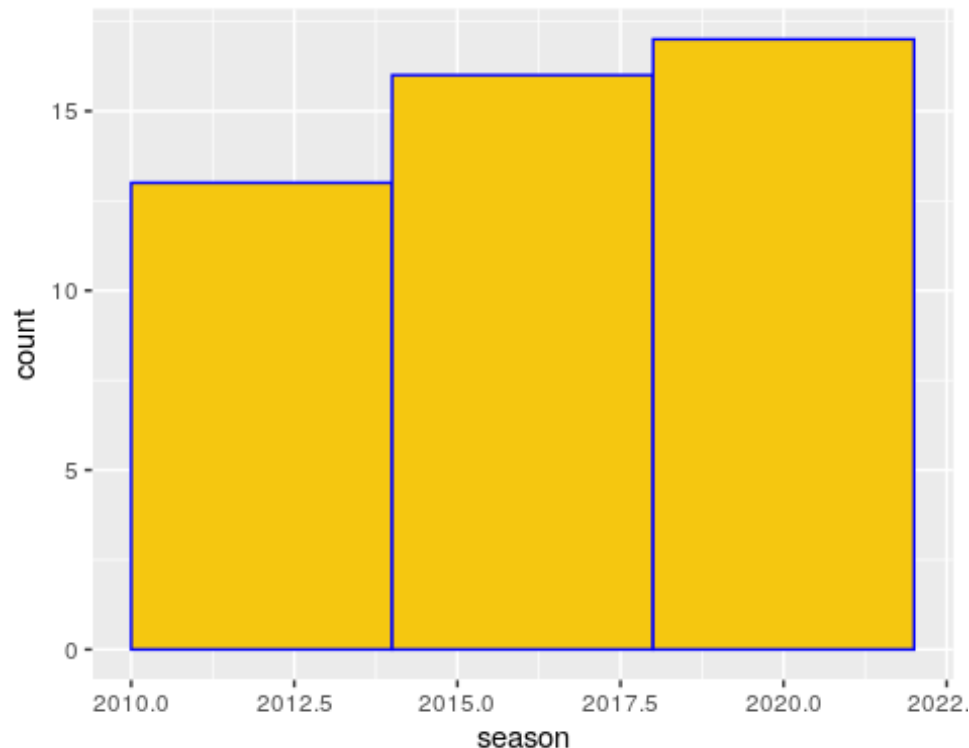
Not surprised!

Calvin Johnson did lead all WRs in TDs for the 2015 season. Second, was Golden Tate, one of the best WRs to break tackles after the catch. I also wanted to see what season did the group of Lions RBs scored the most TDs.

```
player_rb_tds <- nfl %>%group_by(rushing_tds,team,season)
%>%filter(team=="DET")%>%summarize(numofage = n())

## `summarise()` has grouped output by 'rushing_tds', 'team'. You can
## override
## using the `.groups` argument.

ggplot(player_rb_tds,aes(x=season))+
  geom_histogram(binwidth = 4,color="blue",fill = "5647")
```



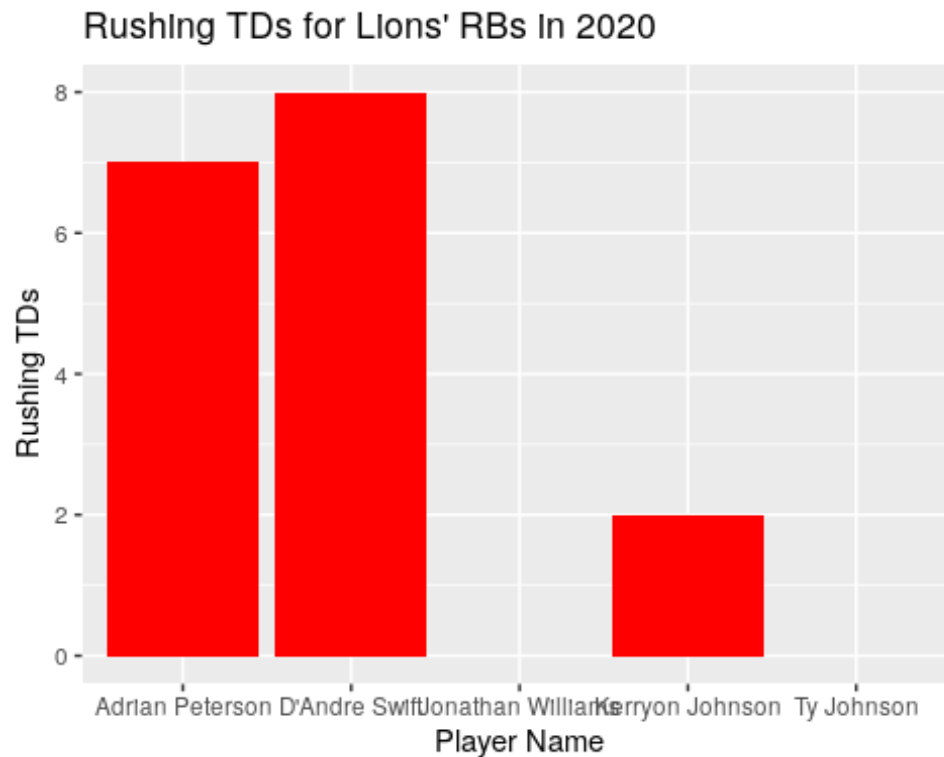
Here, we can see the Lions rushed for more TDs around the 2020 season, totally close to 18 rushing TDs. It looks like the Lions wanted to lean more on the run game than the pass game in 2020. In 2021, the Lions hired new head coach, Dan Campbell, who has established a well balanced offensive game, but has relied more on the rushing game than the passing game. Just like the WR, I wanted to see which player scored the most rushing TDs in the 2020 season. My guess, D'Andre Swift.

```
lions_rb <- nfl%>%group_by(rushing_tds, name, position,team,season)
%>%filter(team=="DET",position=="RB",season=="2020")%>%summarize(numofage =
n())

## `summarise()` has grouped output by 'rushing_tds', 'name', 'position',
'team'.
## You can override using the `.groups` argument.

ggplot(lions_rb, aes(x = name, y = rushing_tds)) +
  geom_histogram(stat = "identity", fill = "red") +
  labs(x = "Player Name", y = "Rushing TDs", title = "Rushing TDs for Lions'
RBs in 2020")

## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



And what do you know, it's D'Andre Swift who led with 8 total TDs! Lastly, I will create a linear model of the weight and 40 yard dash time for all NFL team's WRs, RBs, & TEs. QBs typically don't have impressive 40 yard dash time. Below is the summary table.

```
nfl_teams <- nfl %>%
  filter(position %in% c("WR", "RB", "TE"),
         forty != "N/A",
         wt != "N/A")
nfl_teams <- na.omit(nfl_teams)
nfl_teams$forty <- as.numeric(nfl_teams$forty)
nfl_teams$wt <- as.numeric(nfl_teams$wt)
head(nfl_teams$forty)

## [1] 4.49 4.49 4.49 4.48 4.48 4.48
```

I also added a regression table using the same data I used for the summary table and graphed that data below to show the distribution between weight and the 40 yard dash.

```
regression_model <- lm(forty ~ wt + position, data = nfl_teams)
summary(regression_model)

##
## Call:
## lm(formula = forty ~ wt + position, data = nfl_teams)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32692 -0.06554  0.00080  0.06594  0.30109
```

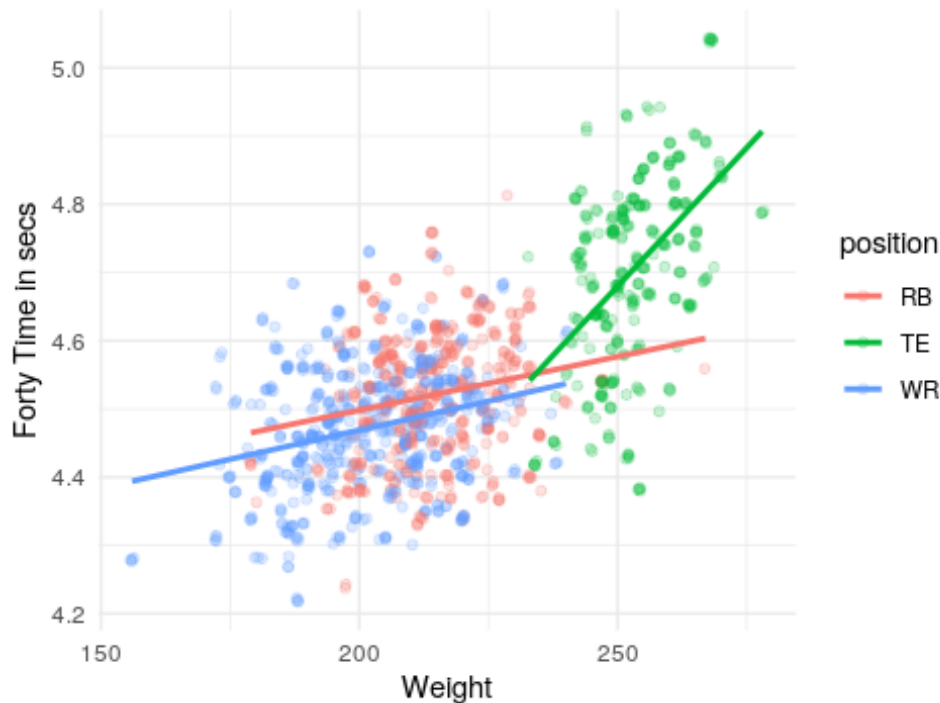
```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0307419   0.0446221   90.331 < 2e-16 ***
## wt          0.0022854   0.0002071   11.033 < 2e-16 ***
## positionTE   0.0956771   0.0107057    8.937 < 2e-16 ***
## positionWR  -0.0201423   0.0063834   -3.155 0.00163 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09922 on 1490 degrees of freedom
## Multiple R-squared:  0.4731, Adjusted R-squared:  0.472
## F-statistic: 445.9 on 3 and 1490 DF, p-value: < 2.2e-16
```

We can see that our median residual is around 0.0008. The estimated coefficient for "wt" is approximately 0.0022854. This indicates the expected change in the response variable (in this case, "forty") per unit increase in the predictor variable "wt," holding other variables constant. Similarly, the estimated coefficients for "positionTE" and "positionWR" are approximately 0.0956771 and -0.0201423. In this case, all the coefficients have p-values that are much smaller than 0.05, indicating that they are statistically significant. Our "intercept", represents the value of the response variable ("forty") when all the predictor variables ("wt," "positionTE," and "positionWR") are zero. In this case, the intercept is approximately 4.0307419. This means that when all the predictor variables are zero, the expected value of "forty" is around 4.03 units.

```
nfl_teams %>%
  ggplot(aes(x = wt, y = forty, color = position)) +
  geom_point(alpha = 0.2, position = "jitter") +
  geom_smooth(method = "lm", se = FALSE, aes(group = position)) +
  theme_minimal() +
  labs(title = "Distribution of Weight and 40 Yards Dash", y="Forty Time in
secs", x="Weight")

## `geom_smooth()` using formula 'y ~ x'
```

Distribution of Weight and 40 Yards Dash



As expected for all NFL team, the TEs would be the heaviest and have the slowest 40 yard dash time in the NFL Combine. The NFL Combine is where the NFL tests college athletes vertical jump, bench press numbers (strength), 40 yard dash (speed) and more! One thing to keep in mind here is not all college football athletes are invited to the NFL Combine. Based on the graph for all NFL teams, the heavier players are, the slower their 40 yard dash will be, which makes sense. One thing that did catch my attention was how WR & RB around 225 pounds, run close to a 4.50 or 4.55 sec. 40 yard dash. While that time is impressive, most NFL scouts like to see for WRs & RBs run a 4.48 sec. It's really interesting to see how only one tenth of a second could make NFL teams be interested or not in a college athlete. But 40 yard dash and time isn't always an indicator for how great players are. Take Devante Adams, WR for the Las Vegas Raiders for example. He ran a 4.56 sec. in the 40 yard dash and is considered one of these best NFL WRs and will be in the Hall of Fame discussion at the end of his career. Or Alvin Kamara, RB for the New Orleans Saints, also ran a 4.56 sec. in the 40 yard dash and is one of the best RBs in the NFL currently. As a prediction, I think NFL scouts shouldn't judge NFL players solely on their speed or weight. Because we have certain super star players who aren't the fastest players but have their own way of playing the game that's making them successful. I believe these results can be promoted and not discourage any college athlete who has a shot to play in the NFL