# NFL Positional

Ian Bogley

9/1/2020

Lets start by loading all of our packages at once.

```
library(pacman)
p_load(rvest,tidyverse,data.table,janitor,cowplot)
```

Now let's scrape some data on NFL salaries for 2019, which we will combine with win-loss data from wikipedia. Let's focus on

```
#exp = expenditures (portion of salary caps on each position)
source_exp_19 <- read_html("https://www.spotrac.com/nfl/positional/breakdown/2019/") %>%
  html_nodes("#main > div > table") %>%
  html_table(fill = TRUE)

exp_19 <- source_exp_19[[1]] %>%
  mutate(
    team = gsub(" ","",Team),
    Team = NULL
    )

exp_19[2:11] <- lapply(exp_19[2:11],function(x) {as.numeric(gsub("\\s([0-9]+).[0-9]M","",x))})

pos_exp_19 <- exp_19 %>% pivot_longer(!c(team, Players)) %>%
  rename(position = name, salary = value)

source_win_19 <- read_html("https://en.wikipedia.org/wiki/2019_NFL_season") %>%
  html_nodes("#mw-content-text > div.mw-parser-output > div:nth-child(68) > table") %>%
  html_table(fill = TRUE)

#win = win loss statistics
win_19 <- source_win_19[[1]] %>%
  select(1:10) %>%
  filter(!grepl("^AFC|NFC",X2),!duplicated(.)) %>%
  row_to_names(row_number = 1) %>%
  rename(team = viewtalkedit) %>%
  mutate(
    PCT = as.numeric(PCT),
    PD = as.integer(PF) - as.integer(PA)
    )
win_19[c(2:4,8,9)] <- lapply(win_19[c(2:4,8,9)],as.integer)
win_19$team <- gsub("[^[:alpha:]]","",win_19$team)

naruki_is_santa <- left_join(pos_exp_19,win_19)
```

```
## Joining, by = "team"
```

```
naruki_is_santa
```

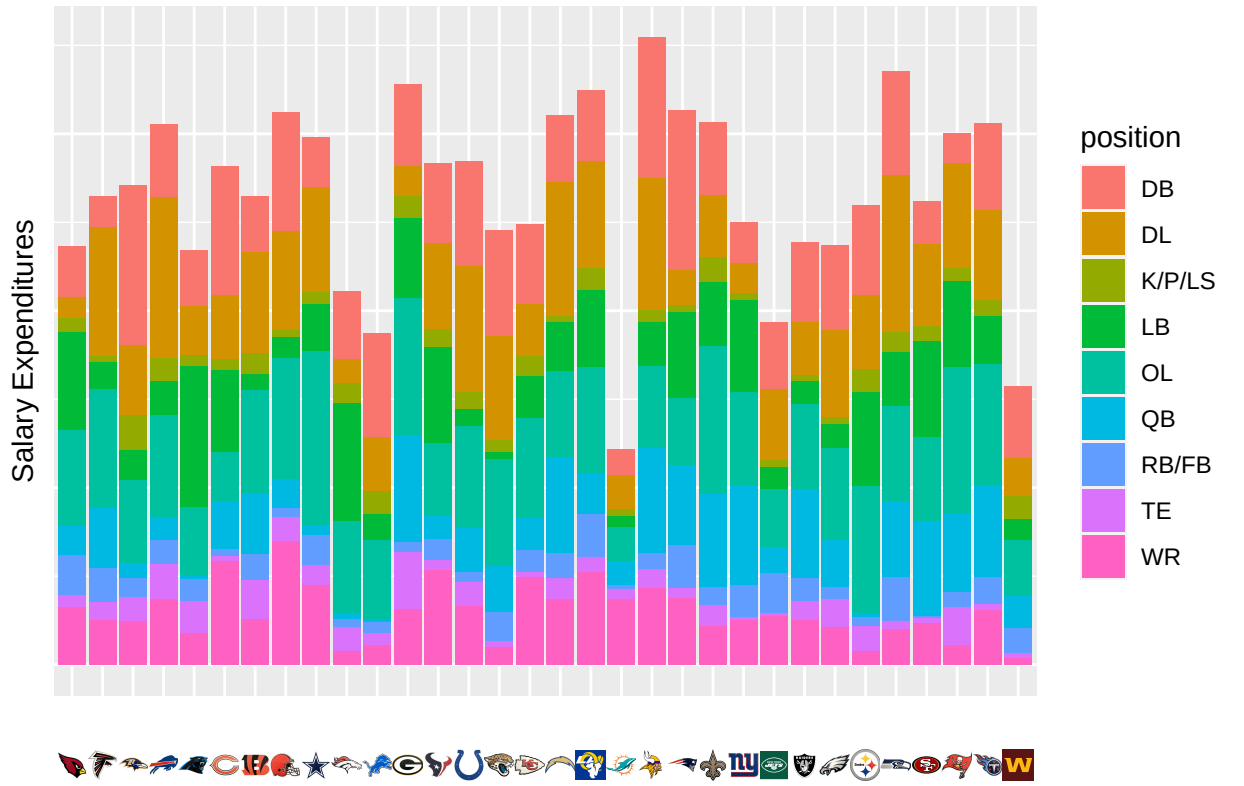
```
## # A tibble: 330 x 14
##    Players team  position salary     W     L     T   PCT DIV   CONF     PF    PA
##      <int> <chr> <chr>     <dbl> <int> <int> <int> <dbl> <chr> <chr> <int> <int>
##  1      53 Ariz~ QB       8.30e6     5    10     1 0.344 1-5   3-8-1   361   442
##  2      53 Ariz~ RB/FB    1.14e7     5    10     1 0.344 1-5   3-8-1   361   442
##  3      53 Ariz~ WR       1.63e7     5    10     1 0.344 1-5   3-8-1   361   442
##  4      53 Ariz~ TE       3.16e6     5    10     1 0.344 1-5   3-8-1   361   442
##  5      53 Ariz~ OL       2.71e7     5    10     1 0.344 1-5   3-8-1   361   442
##  6      53 Ariz~ DL       6.31e6     5    10     1 0.344 1-5   3-8-1   361   442
##  7      53 Ariz~ LB       2.78e7     5    10     1 0.344 1-5   3-8-1   361   442
##  8      53 Ariz~ DB       1.42e7     5    10     1 0.344 1-5   3-8-1   361   442
##  9      53 Ariz~ K/P/LS   3.69e6     5    10     1 0.344 1-5   3-8-1   361   442
## 10      53 Ariz~ Total    1.18e8     5    10     1 0.344 1-5   3-8-1   361   442
## # ... with 320 more rows, and 2 more variables: STK <chr>, PD <int>
```

Next, lets get a barchart of positional data for each team. To start, p will be our base plot.

```
#Graph teams based on salary distributions
p <- naruki_is_santa %>%
  filter(!position == "Total",!team == "Average") %>%
  ggplot() +
  geom_bar(
    aes(
      x = team,
      y = salary,
      fill = position
    ),
    stat = "identity"
  ) +
  theme(
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    axis.title.x = element_blank()
    ) +
  ylab("Salary Expenditures")
```

Next, we will create an x-axis depicting each teams logo. This is implemented through a brute force google image search for each team.

```
ggdraw(insert_xaxis_grob(p,pimage,position = "bottom"))
```




This graph is conclusive evidence that the Miami Dolphins tanked. Ripperino to season pass holders.

However, after some testing I've failed to find significant evidence of significant correlation between positional spending of any kind and win-loss percentage. Also, some light literature review seems to infer that Offense and Defense are each generally equivalent in their effect on team success, implying that differences in positional spending effectiveness might depend on the individual team's approach to offense or defense.

Let's try to use point differential instead, starting with a linear model between point differential and win-loss percentage.

```
lm1 <- naruki_is_santa %>%
  lm(formula = PCT ~ PD)
summary(lm1)
```

```
##
## Call:
## lm(formula = PCT ~ PD, data = .)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.178750 -0.048893 -0.006749  0.038577  0.213842
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 4.989e-01  5.416e-03   92.11   <2e-16 ***
## PD          1.592e-03  5.241e-05   30.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09522 on 308 degrees of freedom
##   (20 observations deleted due to missingness)
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 922.4 on 1 and 308 DF,  p-value: < 2.2e-16
```

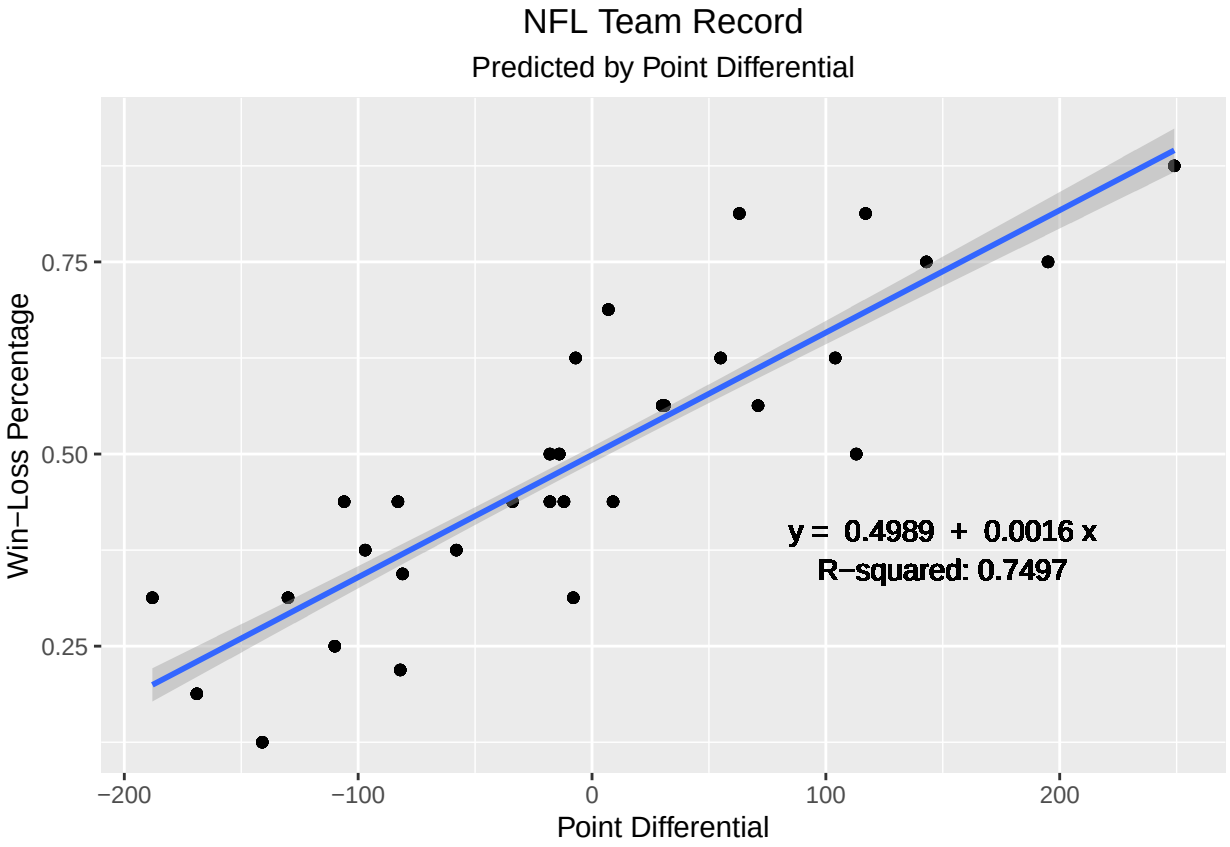
We can see that the relationship is significant, and explains approximately 75% of the variation in win loss percentage.

```
naruki_is_santa %>%
  ggplot(aes(x = PD, y = PCT)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Win-Loss Percentage") +
  xlab("Point Differential") +
  labs(title = "NFL Team Record", subtitle = "Predicted by Point Differential") +
  theme(
    plot.title = element_text(hjust = .5),
    plot.subtitle = element_text(hjust = .5)
    )  +
  geom_text(
    aes(150,.4),
    label = paste("y = ",round(lm1$coefficients[1],digits = 4)," + ",round(lm1$coefficients[2],digits =
  ) +
  geom_text(
    aes(150,.35),
    label = paste("R-squared:",round(summary(lm1)$r.squared,digits = 4))
  )
```

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

```
## Warning: Removed 20 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```


NFL Team Record
Predicted by Point Differential
Win−Loss Percentage
0.75
0.50
0.25
−200
−100
0
100
200
Point Differential
y = 0.4989 + 0.0016 x
R−squared: 0.7497