NFL Statistics and How They Correlate With Yardage

Levi Robertson, Aaryan Thacker, Jonathan Irwanto, Fischer Harrison

```r
#Stats

library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.5.2
```

```r
library(ggrepel)
library(readxl)
NFL_Stats <- read_excel("data/NFL Stats.xlsx")
```

```
New names:
* `` -> `...10`
```

Introduction and Data:

This project is an analysis of different NFL statistics and how they relate to points scored per game. Originally, we compared 3 offensive and 3 defensive statistics to wins (including points and points allowed), which obviously showed that wins were most correlated to points scored. After compiling our data and presenting it, we realized that this project said nothing about who wins in the NFL regular season. Because of this, we changed our focus to comparing 3 offensive statistics (rushing yards, passing yards, turnovers lost) and 3 defensive statistics (rushing yards allowed, passing yards allowed, takeaways) to points per game and points allowed per game respectively. We will attempt to refute the hypothesis that any of these 6 particular statistics do not significantly contribute to the overall points per game (offense) or points allowed per game (defense) statistics.

Methodology:

Ho: The {offensive statistic measured} doesn't statistically contribute to the amount points that a team obtains in any given NFL regular season game from 2014 - 2024

Ha: The {offensive statistic measured} statistically contributes to the amount of points that a team obtains in any given NFL regular season game game from 2014 - 2024

Ho2: The {Defensive statistic measured} doesn't statistically contribute to the amount points that a team allows in any given NFL regular season game from 2014 - 2024

Ha2: The {Defensive statistic measured} statistically contributes to the amount of points that a team allows in any given NFL regular season game from 2014 - 2024

We used a random number generator to make sure this generalized.

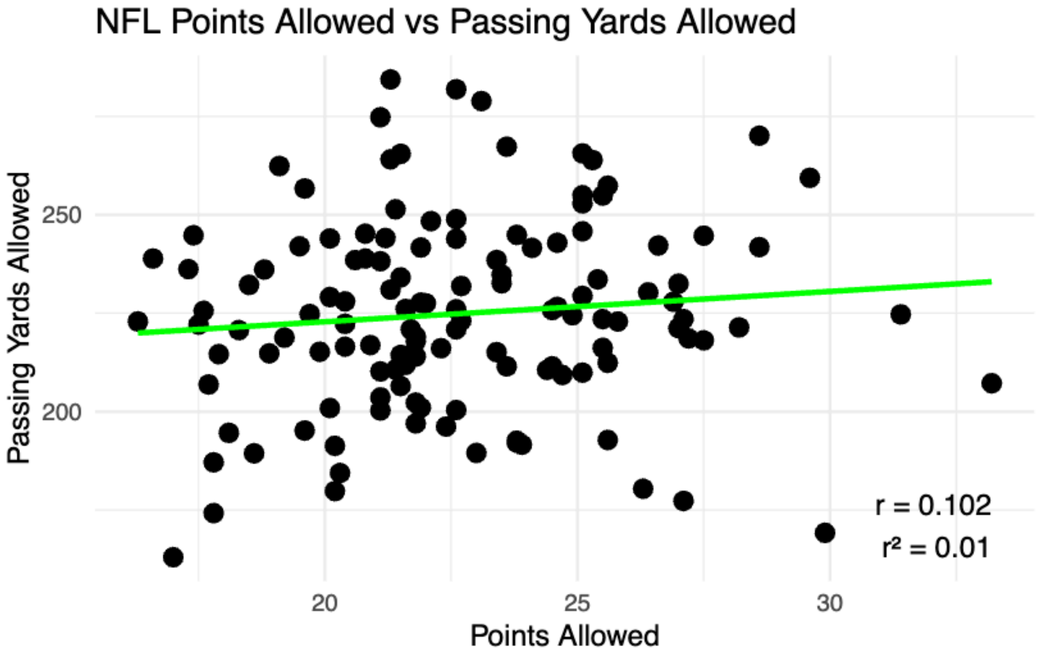
First we grabbed and cleaned our data from pro football reference into R. To test our null and alternative hypothesis we are using the regression test to determine if any are statistically significant. We will be using a confidence interval of 95% percent. Also, we will see which is one the most statistically significant overall for each side of the ball

Results:

Our test informed us that for offense every data point was statistically significant besides TO Lost. The highest R value was passing yards per game with a R value of .511 which shows it has a moderate correlation with wins. For defense we found that only rushing yards allowed was statistically significant and it also had the highest R value. This outcome rejects our null hypothesis because more than one stat on each side of the ball was statistically significant

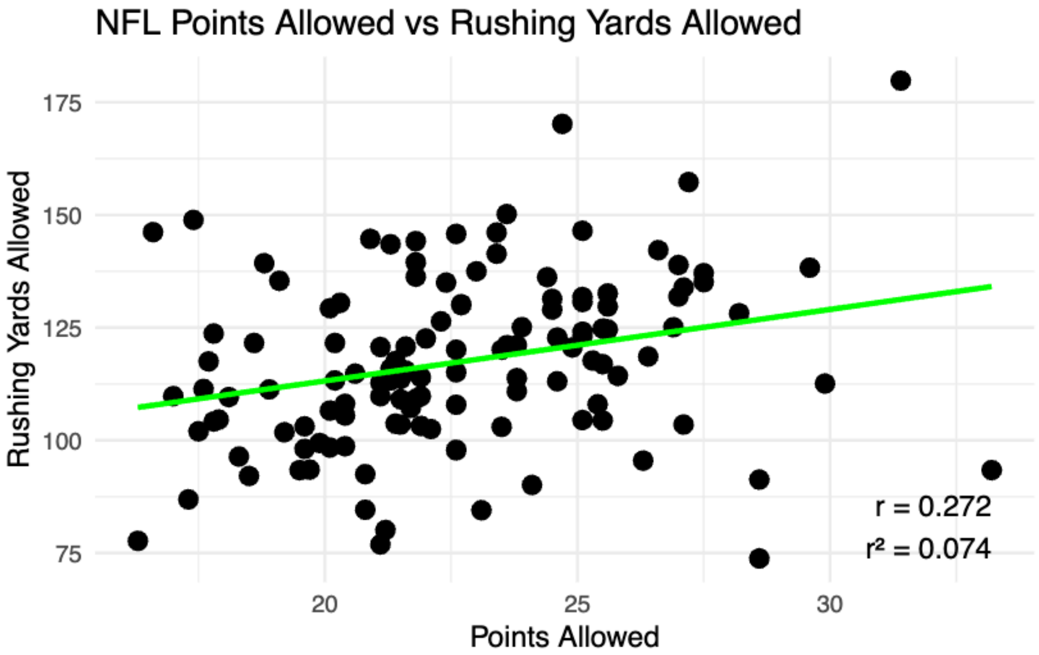
```r
# Helper function to compute r and r2 label
make_r_label <- function(x, y) {
  r <- cor(x, y)
  r2 <- r^2
  paste0("r = ", round(r, 3),
         "\nr² = ", round(r2, 3))
}

### 1. PA vs PYdsA
ggplot(NFL_Stats, aes(x = PA, y = PYdsA, label = '')) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  geom_smooth(method = "lm", color = "green", se = FALSE) +
  labs(x = "Points Allowed", y = "Passing Yards Allowed", title = "NFL Points Allowed vs Pas
  annotate("text",
           x = max(NFL_Stats$PA),
           y = min(NFL_Stats$PYdsA),
           label = make_r_label(NFL_Stats$PA, NFL_Stats$PYdsA),
           hjust = 1, vjust = 0) +
  theme_minimal()
```

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



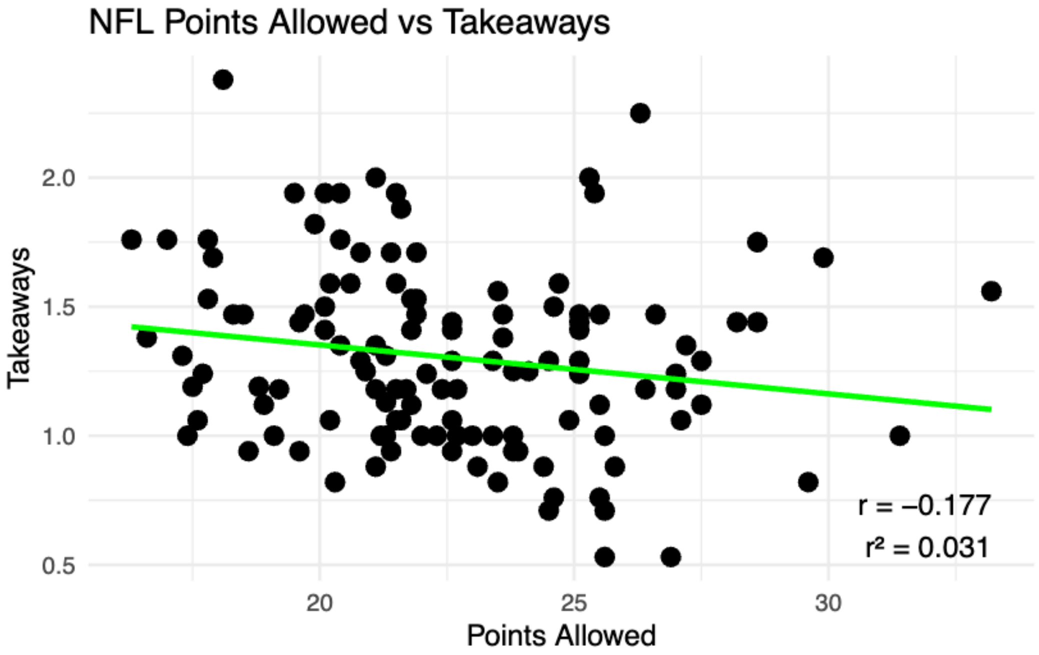
```
## 2. PA vs RYdsA
ggplot(NFL_Stats, aes(x = PA, y = RYdsA, label = '')) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  geom_smooth(method = "lm", color = "green", se = FALSE) +
  labs(x = "Points Allowed", y = "Rushing Yards Allowed", title = "NFL Points Allowed vs Rush
  annotate("text",
           x = max(NFL_Stats$PA),
           y = min(NFL_Stats$RYdsA),
           label = make_r_label(NFL_Stats$PA, NFL_Stats$RYdsA),
           hjust = 1, vjust = 0) +
  theme_minimal()
```

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



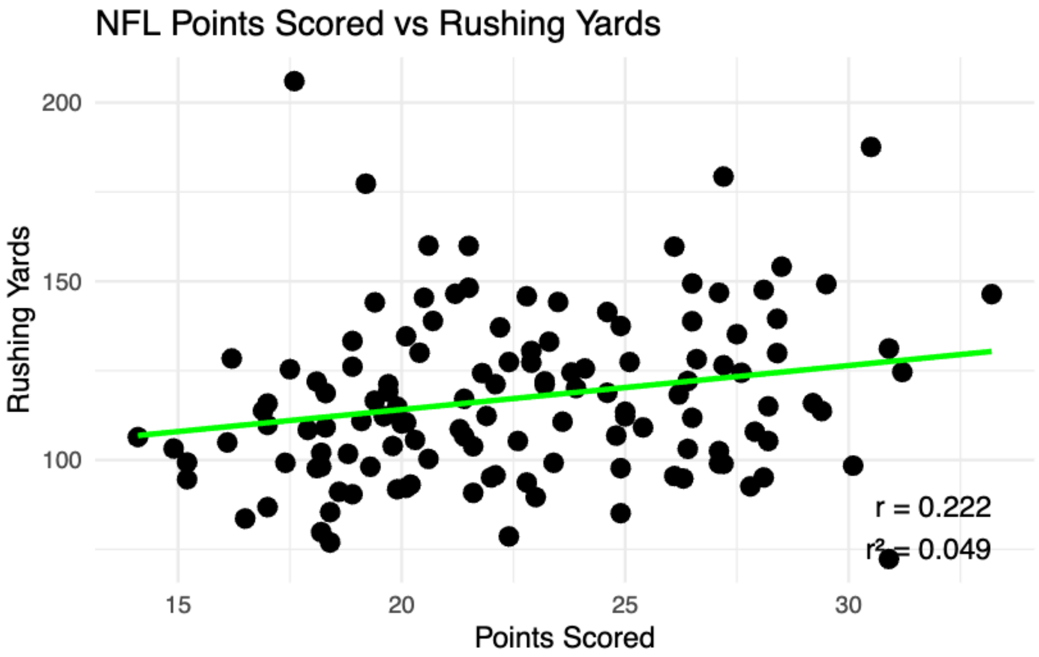
```r
##PA vs Takeaways
ggplot(NFL_Stats, aes(x = PA, y = `Takeaways(Defense)`, label = '')) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  geom_smooth(method = "lm", color = "green", se = FALSE) +
  labs(x = "Points Allowed", y = "Takeaways", title = "NFL Points Allowed vs Takeaways") +
  annotate("text",
           x = max(NFL_Stats$PA),
           y = min(NFL_Stats$`Takeaways(Defense)`),
           label = make_r_label(NFL_Stats$PA, NFL_Stats$`Takeaways(Defense)`),
           hjust = 1, vjust = 0) +
  theme_minimal()
```

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



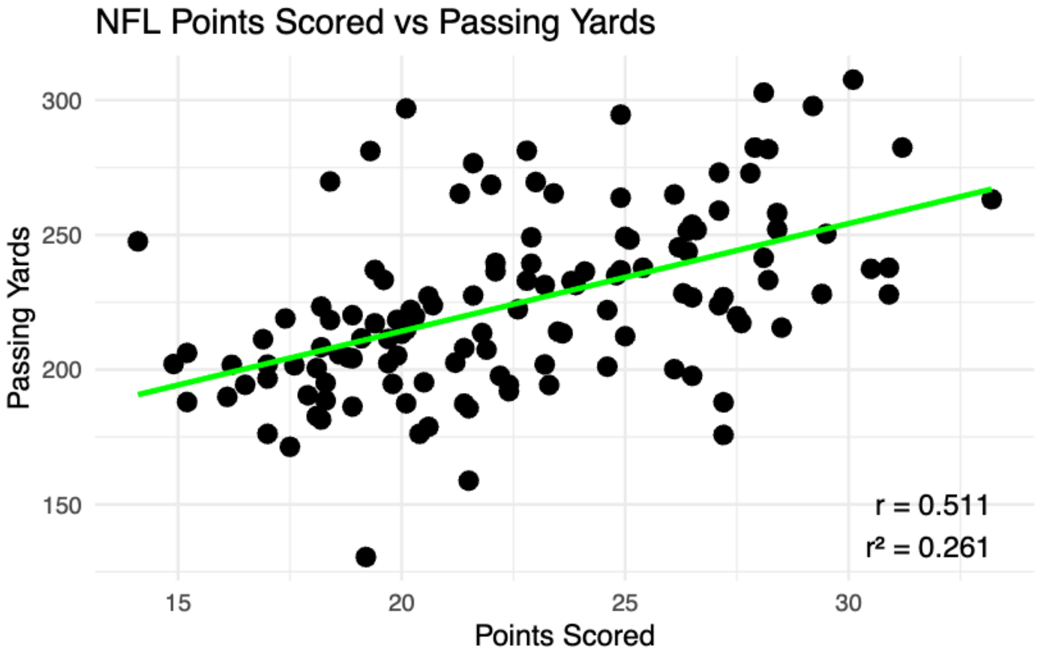
```
##Points For vs RY/G
ggplot(NFL_Stats, aes(x = `Pts For`, y = `RY/G`, label = '')) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  geom_smooth(method = "lm", color = "green", se = FALSE) +
  labs(x = "Points Scored", y = "Rushing Yards", title = "NFL Points Scored vs Rushing Yards"
  annotate("text",
           x = max(NFL_Stats$`Pts For`),
           y = min(NFL_Stats$`RY/G`),
           label = make_r_label(NFL_Stats$`Pts For`, NFL_Stats$`RY/G`),
           hjust = 1, vjust = 0) +
  theme_minimal()
```

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



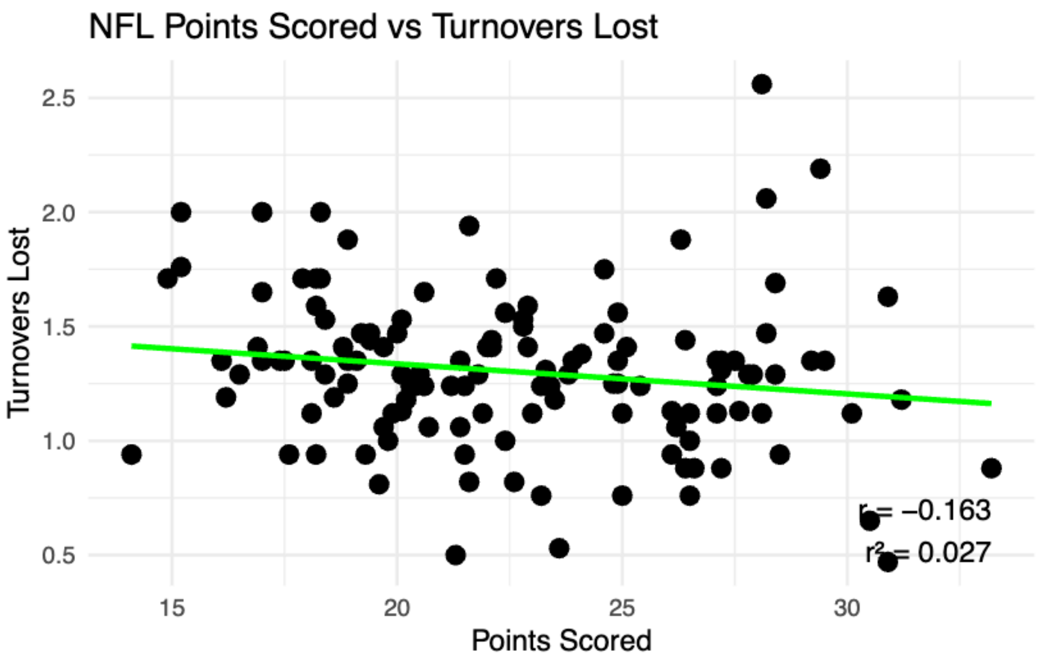
```r
## Pts For vs PY/G
ggplot(NFL_Stats, aes(x = `Pts For`, y = `PY/G`, label = '')) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  geom_smooth(method = "lm", color = "green", se = FALSE) +
  labs(x = "Points Scored", y = "Passing Yards", title = "NFL Points Scored vs Passing Yards"
  annotate("text",
           x = max(NFL_Stats$`Pts For`),
           y = min(NFL_Stats$`PY/G`),
           label = make_r_label(NFL_Stats$`Pts For`, NFL_Stats$`PY/G`),
           hjust = 1, vjust = 0) +
  theme_minimal()
```

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



```
## Pts For vs TO Lost
ggplot(NFL_Stats, aes(x = `Pts For`, y = `TO Lost (Offense)`, label = '')) +
  geom_point(size = 3) +
  geom_text_repel(size = 3) +
  geom_smooth(method = "lm", color = "green", se = FALSE) +
  labs(x = "Points Scored", y = "Turnovers Lost", title = "NFL Points Scored vs Turnovers Los
  annotate("text",
           x = max(NFL_Stats$`Pts For`),
           y = min(NFL_Stats$`TO Lost (Offense)`),
           label = make_r_label(NFL_Stats$`Pts For`, NFL_Stats$`TO Lost (Offense)`),
           hjust = 1, vjust = 0) +
  theme_minimal()
```

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



```
#p-value
options(scipen = 5)

offense_model <- lm(`Pts For` ~ `PY/G` + `RY/G` + `TO Lost (Offense)`,
                    data =NFL_Stats)

defense_model <- lm(PA ~ PYdsA + RYdsA + `Takeaways(Defense)`,
                    data = NFL_Stats)

offense_coefs <- summary(offense_model)$coefficients
offense_coefs
```

```
                       Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)         -1.98125035 3.510528265 -0.5643739 5.735192e-01
`PY/G`               0.07738038 0.009301797  8.3188641 1.350636e-13
`RY/G`               0.06611489 0.013874071  4.7653562 5.178905e-06
`TO Lost (Offense)` -0.40312039 0.919522143 -0.4384020 6.618577e-01
```

```
defense_coefs <- summary(defense_model)$coefficients
defense_coefs
```

```
                        Estimate Std. Error   t value     Pr(>|t|)
```

```
(Intercept)          14.06067090 3.89228510  3.612446 0.0004390431
PYdsA                 0.01877278 0.01159440  1.619125 0.1079611270
RYdsA                 0.04711618 0.01535348  3.068762 0.0026405964
`Takeaways(Defense)` -0.89958022 0.82562683 -1.089572 0.2780144678
```

Conclusions and Future Work:

By performing this statistical analysis we have come to the conclusion that offensive statistics are much more indicative of points per game than defensive stats are of points allowed per game. Because the p-values obtained from each offensive statistic (excluding takeaways) are much smaller than the p-values obtained from each defensive statistic (though rushing yards allowed per game is still statistically significant), the data suggests that the strength of the offense of any given NFL team during the regular seasons from 2019-2024 is a much stronger deciding factor in points scored/allowed than the strength of the defense of that same team. Our analysis was limited to the regular season because post-season data was difficult to compile, and would also conflict with the teams that do not get post-season play. This analysis could be improved with a larger sampling pool, although going too far back may change the trends because NFL playcalling changes as time goes on. Our data also does not account for injuries. In the future, we could investigate potential trade outlooks for different teams, adding to our scope star players for each team and compare results to see which team needs players who have high scores in statistics that the team is lacking in.