Data Mining Project

Pro Football Reference

Ben Winkler, Gabriel Gutierrez, Paul Rivera, & Siena Adams

Executive Summary

Overview

- This report uses real-time 2025 NFL season statistics to determine which teams have the strongest profiles (Up to Week 13)
- Data collected:
    - Team-level offensive
    - Team-level defensive
    - Advanced Defense
    - Drive Outcomes
        - Outcomes
            - Points scored
            - Points allowed
            - Total yards gained
            - Yards allowed
            - Turnovers
            - Takeaways
- These variables allow us to analyze performance patterns across the league

Uses

Descriptive:

- Summarize & visualize how teams across the NFL are performing through the 2025 season

Prescriptive:

- Provide actionable recommendations for how teams could increase their chances of winning

Predictive:

- Build a statistical model that predicts a team's likelihood of future success

Cleaning the Data

First Step:

- Fixed issue with column headers, removing primary heading to make way for more specific field names.
- Renamed column headers to specify total statistics, passing statistics, and rushing statistics with the same name (Yards, TD, Attempts, 1st Down)
    - Both of these steps were done within the CSV file

Second Step:

- Removed columns relating to penalties, which we do not anticipate will provide significant predictive, descriptive or prescriptive value. (Less penalties = good)
- Removed last 3 observations for both Team Offense and Defense, which contained league averages and league totals (avoid contaminating regression)
- Removed last observation for Drive Averages containing league wide totals
- Removed Percentage character (%) from all columns containing "%" in order to confirm these variables as continuous rather than categorical

```
offense_raw <- offense_raw[, !names(offense_raw) %in% "Pen"]
defense_raw <- defense_raw[, !names(defense_raw) %in% "Pen"]

offense_raw <- offense_raw[, !names(offense_raw) %in% "Yds.1"]
defense_raw <- defense_raw[, !names(defense_raw) %in% "Yds.1"]

offense_raw <- offense_raw[, !names(offense_raw) %in% "X1stPy"]
defense_raw <- defense_raw[, !names(defense_raw) %in% "X1stPy"]

offense_raw <- offense_raw[, !names(offense_raw) %in% "FL"]
defense_raw <- defense_raw[, !names(defense_raw) %in% "FL"]
```

```
offense_raw <- offense_raw[1:(nrow(offense_raw) - 3), ]
defense_raw <- defense_raw[1:(nrow(defense_raw) - 3), ]
```

```
advanced_defense$Bltz. <- as.numeric(gsub("%", "", advanced_defense$Bltz.))
advanced_defense$Prss. <- as.numeric(gsub("%", "", advanced_defense$Prss.))
```

Descriptive Analysis

SQL Query #1

```
#Top 10 teams in Passing Efficiency
library(sqldf)
sqldf("SELECT Tm, `NY.A` FROM offense_raw ORDER BY `NY.A` DESC LIMIT 10")
```

```
> library(sqldf)
> sqldf("SELECT Tm, `NY.A` FROM offense_raw ORDER BY `NY.A` DESC LIMIT 10")
                    Tm NY.A
1      Seattle Seahawks  8.2
2  New England Patriots  7.5
3    Indianapolis Colts  7.2
4      Los Angeles Rams  7.1
5        Dallas Cowboys  7.0
6         Detroit Lions  7.0
7         Buffalo Bills  7.0
8     Green Bay Packers  7.0
9   San Francisco 49ers  6.7
10   Kansas City Chiefs  6.6
```

SQL Query #2

```
#Teams forcing the most interceptions
sqldf("SELECT Tm, `Int` FROM defense_raw ORDER BY `Int` DESC")
```

```
> #Rushing efficiency leaders
> sqldf("SELECT Tm, `Int` FROM defense_raw ORDER BY `Int` DESC")
                      Tm Int
1           Chicago Bears  17
2         Seattle Seahawks  13
3    Jacksonville Jaguars  13
4          Houston Texans  12
5         Los Angeles Rams  12
6       Indianapolis Colts  12
7        Carolina Panthers  12
8    Los Angeles Chargers  11
9    Tampa Bay Buccaneers  11
10    Pittsburgh Steelers  10
11     Cincinnati Bengals  10
12       Cleveland Browns   9
13          Detroit Lions   9
14         Atlanta Falcons   9
15       Arizona Cardinals   9
16  New England Patriots   8
17    Philadelphia Eagles   8
18          Buffalo Bills   8
19       Las Vegas Raiders   8
20          Denver Broncos   7
21      Kansas City Chiefs   7
22        Baltimore Ravens   7
23      New Orleans Saints   7
24       Green Bay Packers   6
25 Washington Commanders   6
26    San Francisco 49ers   5
27       Tennessee Titans   5
28         New York Giants   5
29          Dallas Cowboys   5
30          Miami Dolphins   4
31       Minnesota Vikings   3
32          New York Jets   0
```

SQL Query #3

```
#Teams with the most turnovers
sqldf("SELECT Tm, `TO` FROM offense_raw ORDER BY `TO` DESC")
```

```
> #Teams with the most turnovers
> sqldf("SELECT Tm, `TO` FROM offense_raw ORDER BY `TO` DESC")
                    Tm TO
1      Minnesota Vikings 26
2       Seattle Seahawks 22
3     New Orleans Saints 19
4    San Francisco 49ers 18
5       Baltimore Ravens 18
6     Cincinnati Bengals 18
7         Miami Dolphins 18
8          Buffalo Bills 17
9      Carolina Panthers 17
10     Las Vegas Raiders 17
11      Tennessee Titans 16
12        Dallas Cowboys 15
13     Arizona Cardinals 15
14 Washington Commanders 15
15      Cleveland Browns 15
16  Jacksonville Jaguars 14
17   Pittsburgh Steelers 14
18  Los Angeles Chargers 14
19         New York Jets 14
20  New England Patriots 13
21        Denver Broncos 13
22       New York Giants 13
23    Indianapolis Colts 12
24      Los Angeles Rams 11
25       Atlanta Falcons 11
26        Houston Texans 10
27         Chicago Bears  9
28  Tampa Bay Buccaneers  9
29         Detroit Lions  8
30    Kansas City Chiefs  8
31   Philadelphia Eagles  8
32     Green Bay Packers  7
```

SQL Query #4

```
#Best pass defenses (lowest NY/A allowed)
sqldf("SELECT Tm, `NY.A` FROM defense_raw ORDER BY `NY.A` ASC")
```

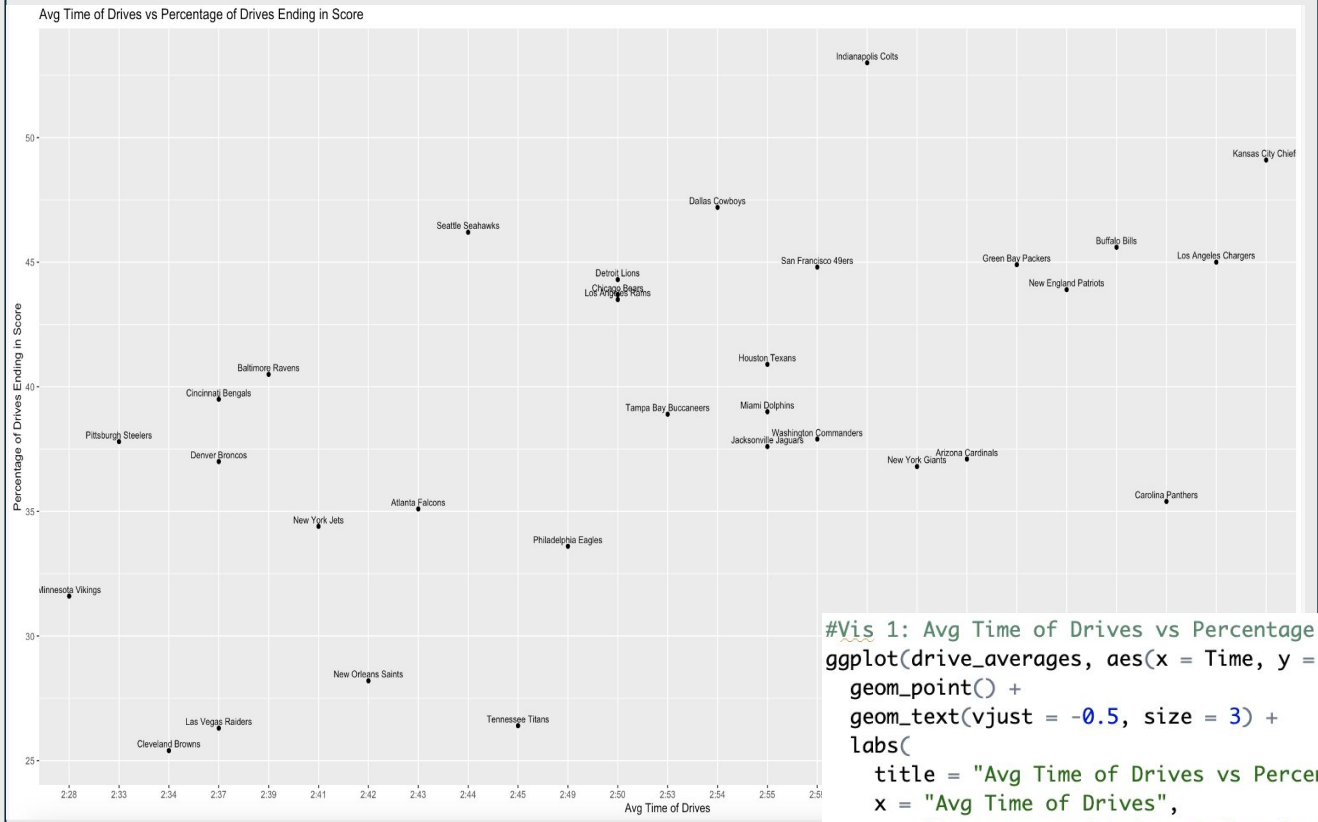
```
> #Best pass defenses (lowest NY/A allowed)
> sqldf("SELECT Tm, `NY.A` FROM defense_raw ORDER BY `NY.A` ASC")
                    Tm NY.A
1         Denver Broncos  4.9
2       Seattle Seahawks  5.0
3         Houston Texans  5.1
4      Green Bay Packers  5.2
5   Los Angeles Chargers  5.3
6       Cleveland Browns  5.4
7       Los Angeles Rams  5.5
8          Buffalo Bills  5.5
9        Atlanta Falcons  5.6
10  Jacksonville Jaguars  5.8
11    Minnesota Vikings  5.9
12    Indianapolis Colts  6.0
13   Philadelphia Eagles  6.0
14  New England Patriots  6.1
15   Pittsburgh Steelers  6.1
16      Baltimore Ravens  6.1
17       New York Giants  6.1
18         Detroit Lions  6.2
19     Arizona Cardinals  6.2
20         New York Jets  6.2
21    New Orleans Saints  6.3
22    Kansas City Chiefs  6.5
23   San Francisco 49ers  6.5
24     Las Vegas Raiders  6.5
25        Miami Dolphins  6.6
26     Carolina Panthers  6.6
27  Tampa Bay Buccaneers  6.7
28      Tennessee Titans  6.7
29         Chicago Bears  6.8
30        Dallas Cowboys  6.8
31    Cincinnati Bengals  7.3
32 Washington Commanders  7.7
```

SQL Query #5

```
#Top blitzing teams
sqldf("SELECT Tm, `Bltz.` FROM advanced_defense ORDER BY `Bltz.` DESC")
```

```
> #Top blitzing teams
> sqldf("SELECT Tm, `Bltz.` FROM advanced_defense ORDER BY `Bltz.` DESC")
                    Tm Bltz.
1      Minnesota Vikings  43.8
2        Atlanta Falcons  35.8
3     Kansas City Chiefs  31.4
4          Miami Dolphins  30.4
5   Tampa Bay Buccaneers  30.4
6    Pittsburgh Steelers  30.1
7       Cleveland Browns  27.5
8          New York Jets  27.5
9          Chicago Bears  26.6
10        Denver Broncos  26.6
11        Dallas Cowboys  25.7
12     Arizona Cardinals  25.5
13    New Orleans Saints  25.5
14      Baltimore Ravens  25.5
15         Detroit Lions  24.4
16  Jacksonville Jaguars  24.4
17       New York Giants  23.5
18        Houston Texans  22.8
19 Washington Commanders  22.8
20  New England Patriots  22.6
21    Indianapolis Colts  22.5
22      Tennessee Titans  21.3
23   Philadelphia Eagles  20.6
24         Buffalo Bills  20.3
25     Las Vegas Raiders  19.9
26      Seattle Seahawks  19.9
27  Los Angeles Chargers  19.0
28     Carolina Panthers  18.9
29      Los Angeles Rams  18.7
30     Green Bay Packers  18.0
31   San Francisco 49ers  17.5
32    Cincinnati Bengals  17.3
```

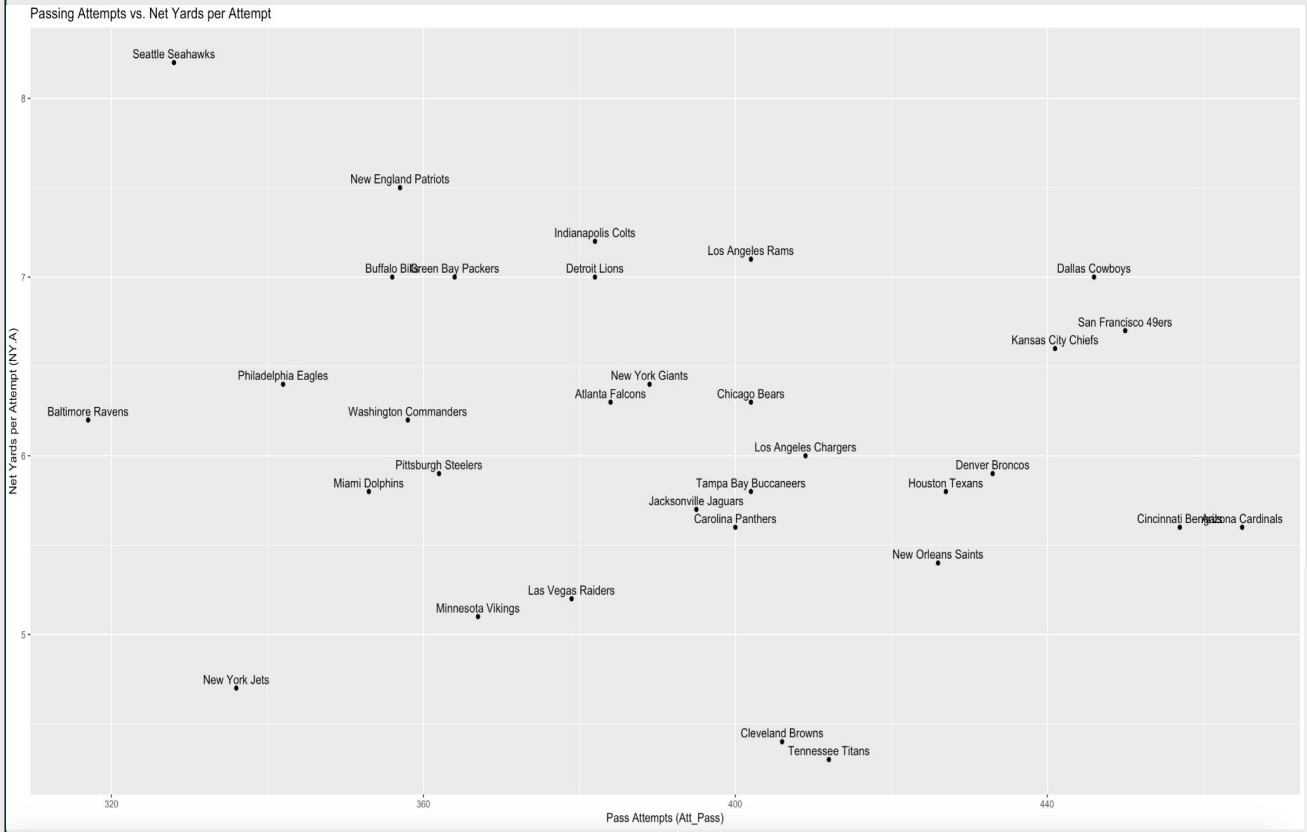
Visualization #1: AVG. Drive Length vs % of Drives Scored (Scatter Plot)



The longer control of the ball, the higher % of the drive ending in a score.

```r
#Vis 1: Avg Time of Drives vs Percentage of Drives Ending in Score (Scatterplot)
ggplot(drive_averages, aes(x = Time, y = Sc., label = Tm)) +
  geom_point() +
  geom_text(vjust = -0.5, size = 3) +
  labs(
    title = "Avg Time of Drives vs Percentage of Drives Ending in Score",
    x = "Avg Time of Drives",
    y = "Percentage of Drives Ending in Score"
  )
```

Visualization #2: Passing Attempts vs Net Yards per Attempt (Scatter Plot)

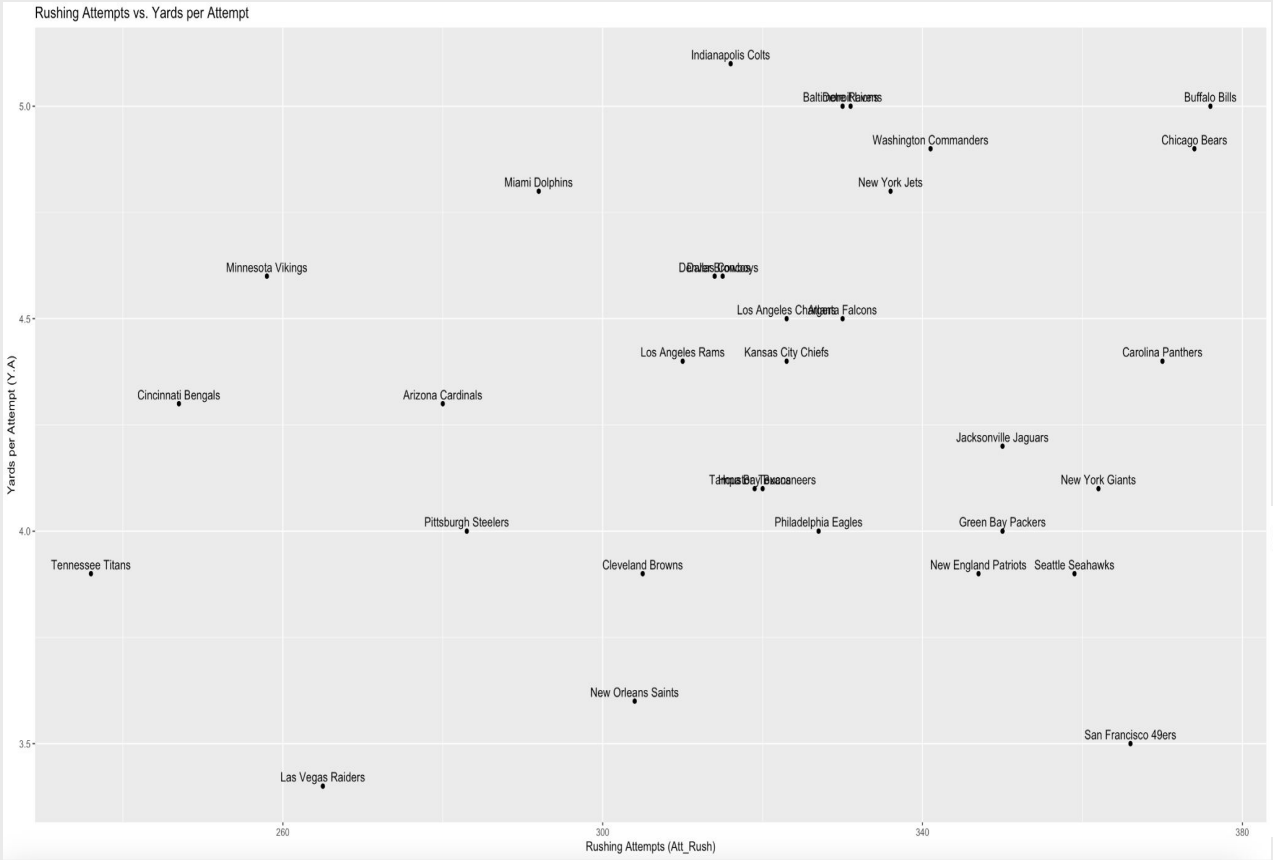


Passing Efficiency

Less pass attempts, higher pass efficiency for offenses

```
ggplot(offense_raw, aes(x = Att_Pass, y = NY.A)) +
  geom_point() +
  geom_text(aes(label = Tm), vjust = -0.5) +
  labs(
    title = "Passing Attempts vs. Net Yards per Attempt",
    x = "Pass Attempts (Att_Pass)",
    y = "Net Yards per Attempt (NY.A)"
  )
```

Visualization #3: Rushing Attempts vs. Yards per carry (Scatter Plot)

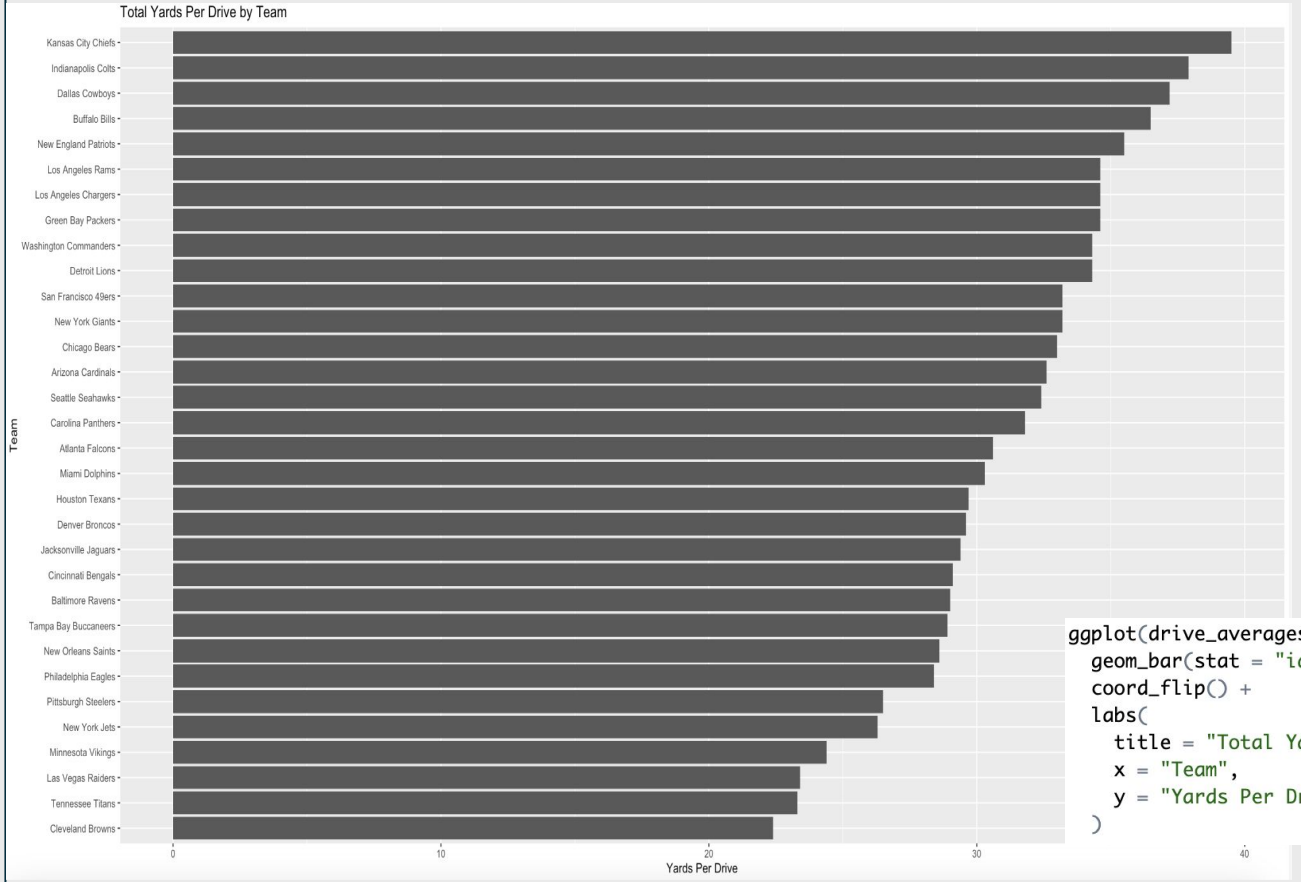


Rushing Efficiency

More rushing attempts, higher rush efficiency for offenses

```
ggplot(offense_raw, aes(x = Att_Rush, y = Y.A)) +
  geom_point() +
  geom_text(aes(label = Tm), vjust = -0.5) +
  labs(
    title = "Rushing Attempts vs. Yards per Attempt",
    x = "Rushing Attempts (Att_Rush)",
    y = "Yards per Attempt (Y.A)"
  )
```

Visualisation #4: Ranking of Yards Per Drive (Horizontal Bar Chart)



Most effective offenses at moving the ball

The better the offense, higher the avg. yards per drive.

```
ggplot(drive_averages, aes(x = reorder(Tm, Yds), y = Yds)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(
    title = "Total Yards Per Drive by Team",
    x = "Team",
    y = "Yards Per Drive"
  )
```

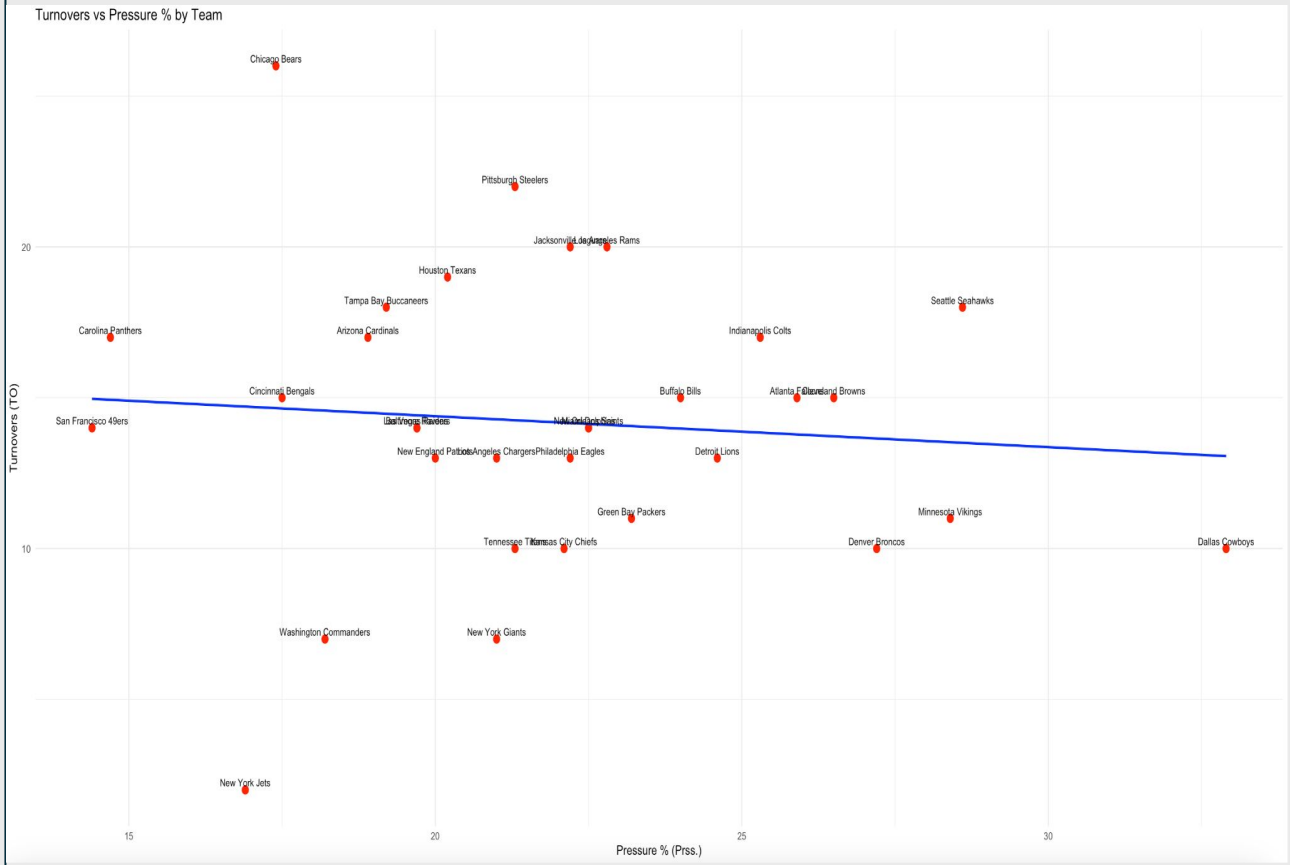
Visualisation #5: Heat Map of Total Yards Allowed by Team (Heat Map)

Overall defensive strength

The darker and bigger the bubble, the worse the overall defense

```
defense_plot <- defense_raw %>%
  arrange(desc(Yds)) %>%
  mutate(Team = factor(Tm, levels = Tm))

ggplot(defense_raw, aes(x = "Yds", y = Tm)) +
  geom_point(aes(size = Yds, color = Yds)) +
  scale_size_continuous(range = c(3, 15)) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(
    title = "Bubble Heatmap of Defensive Yards Allowed",
    x = "",
    y = "Team",
    size = "Yards Allowed",
    color = "Yards Allowed"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
  )
```

Visualisation #6: Advanced_defense Pressure Rate to defense_raw TO% (Scatter Plot)

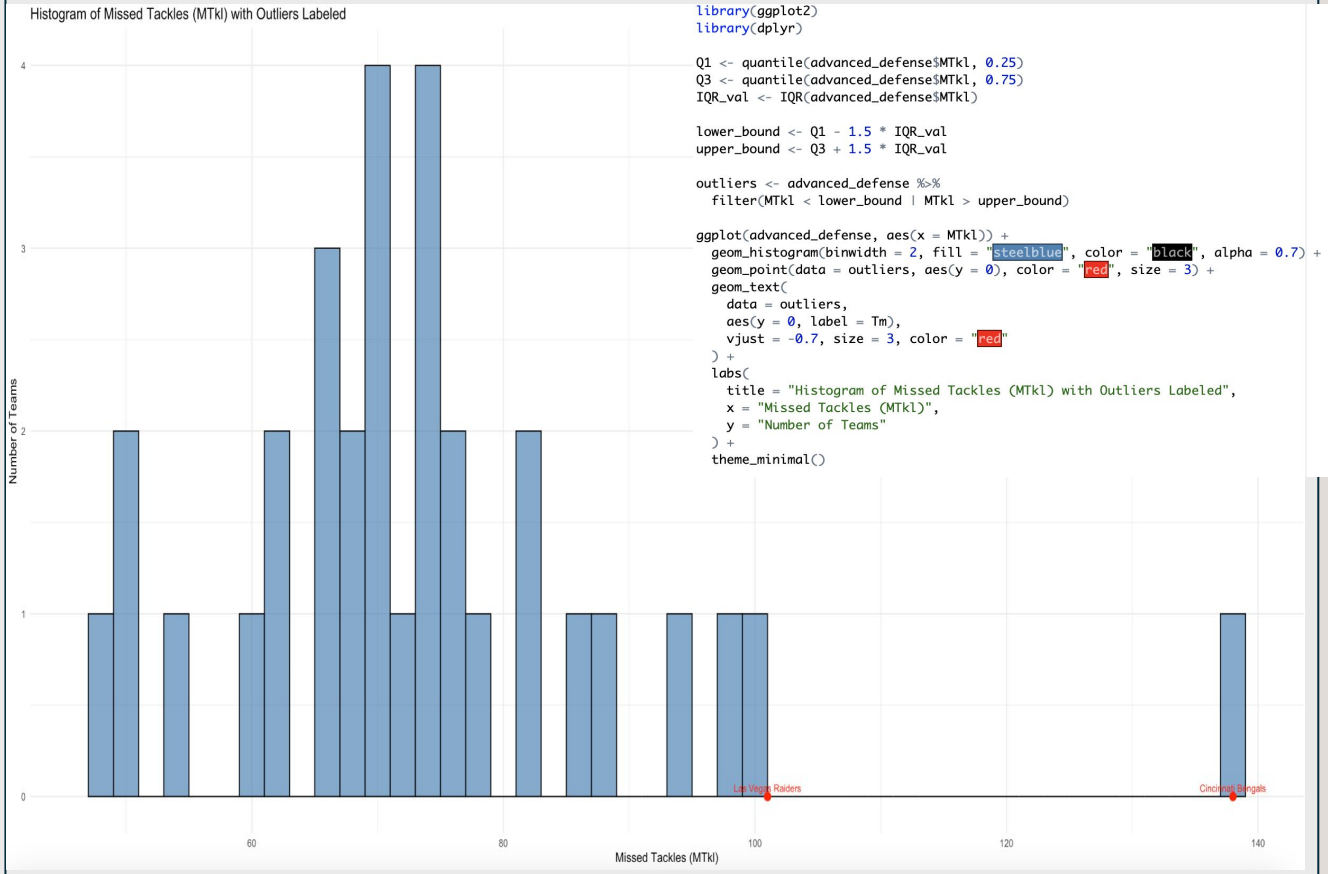


How effectively pressure converts to turnovers

The less the pressure %, the more likely to get a turnover

```r
# 1. Merge the datasets by team
combined <- defense_raw %>%
  inner_join(advanced_defense, by = "Tm")

# 2. Ensure Prss. is numeric (if it has % signs)
combined$Prss. <- as.numeric(gsub("%", "", combined$Prss.))

# Optional: convert to decimal
# combined$Prss. <- combined$Prss. / 100

# 3. Create scatterplot with labels (without ggrepel)
ggplot(combined, aes(x = Prss., y = TO)) +
  geom_point(color = "red", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  geom_text(aes(label = Tm), vjust = -0.5, hjust = 0.5, size = 3) +  # label points
  labs(
    x = "Pressure % (Prss.)",
    y = "Turnovers (TO)",
    title = "Turnovers vs Pressure % by Team"
  ) +
  theme_minimal()
```

Visualisation #7: advanced_defense MTkl (Histogram)

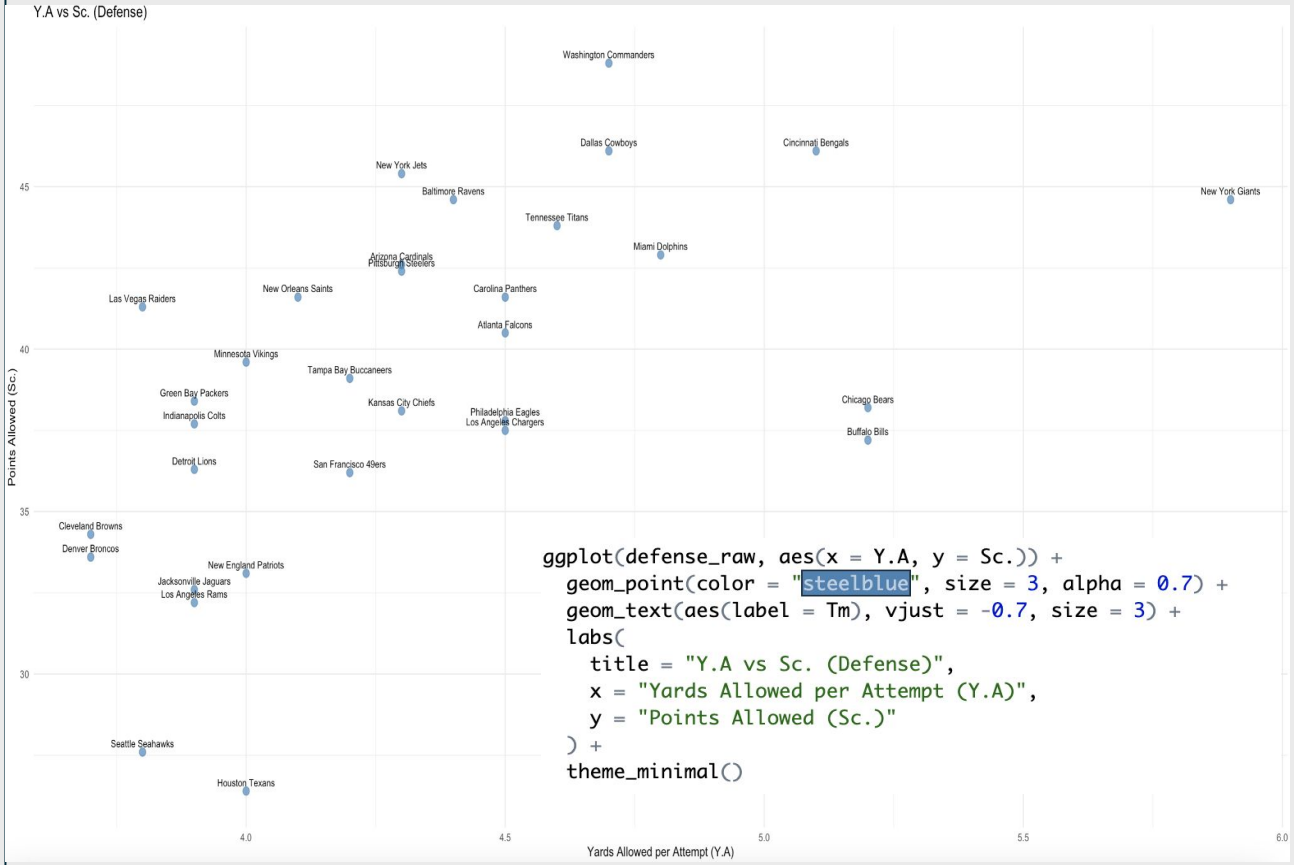



```r
library(ggplot2)
library(dplyr)

Q1 <- quantile(advanced_defense$MTkl, 0.25)
Q3 <- quantile(advanced_defense$MTkl, 0.75)
IQR_val <- IQR(advanced_defense$MTkl)

lower_bound <- Q1 - 1.5 * IQR_val
upper_bound <- Q3 + 1.5 * IQR_val

outliers <- advanced_defense %>%
  filter(MTkl < lower_bound | MTkl > upper_bound)

ggplot(advanced_defense, aes(x = MTkl)) +
  geom_histogram(binwidth = 2, fill = "steelblue", color = "black", alpha = 0.7) +
  geom_point(data = outliers, aes(y = 0), color = "red", size = 3) +
  geom_text(
    data = outliers,
    aes(y = 0, label = Tm),
    vjust = -0.7, size = 3, color = "red"
  ) +
  labs(
    title = "Histogram of Missed Tackles (MTkl) with Outliers Labeled",
    x = "Missed Tackles (MTkl)",
    y = "Number of Teams"
  ) +
  theme_minimal()
```

Teams missing the most tackles

2 outliers when it comes to tackles missed

Visualisation #8: defense_raw Rush Y/A vs defense_raw Sc%




```
ggplot(defense_raw, aes(x = Y.A, y = Sc.)) +
  geom_point(color = "steelblue", size = 3, alpha = 0.7) +
  geom_text(aes(label = Tm), vjust = -0.7, size = 3) +
  labs(
    title = "Y.A vs Sc. (Defense)",
    x = "Yards Allowed per Attempt (Y.A)",
    y = "Points Allowed (Sc.)"
  ) +
  theme_minimal()
```

Rushing defense efficiency compared to overall scoring efficiency allowed

The lower the yards per attempt indicates a strong rushing defense

Predictive Analysis

What are “Expected Points”?

- Composite efficiency metric
- Based on advanced statistics
- More stable and predictive than raw points
- Considered a gold-standard valuation of team strength

Last 5 Super Bowl Winners (EXP Ranks)

2025 - Philadelphia Eagles (6th, 1st)

2024 - Kansas City Chiefs (10th, 4th)

2023 - Kansas City Chiefs (1st, 20th)

2022 - Los Angeles Rams (6th, 8th)

2021 - Tampa Bay Buccaneers (3rd, 6th)

Offensive Regression

```
offense_model <- lm(EXP ~ NY.A + Sc. + TO. + Y.P, data = offense_raw)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -434.452     49.580  -8.763 2.23e-09
NY.A          14.080      9.516   1.480 0.150533
Sc.            3.846      1.107   3.476 0.001739
TO.           -4.470      1.157  -3.865 0.000632
Y.P           53.449     20.054   2.665 0.012829
```

```
Multiple R-squared:  0.9355,    Adjusted R-squared:  0.9259
F-statistic: 97.85 on 4 and 27 DF,  p-value: 1.166e-15
```

1. Net Yards per Attempt (NY/A)
2. Percentage of Drives Ending in Score (Sc%)
3. Turnover Rate (TO%)
4. Yards per Play (Y/P)

Defensive Regression

```
defense_model <- lm(EXP ~ NY.A + Sc. + TO. + Y.P, data = defense_raw)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  438.797     39.558  11.092 1.47e-11
NY.A           9.537     12.154   0.785  0.43946
Sc.           -3.734      1.116  -3.344  0.00243
TO.            2.206      1.091   2.022  0.05318
Y.P          -77.771     16.653  -4.670 7.39e-05
```

```
Multiple R-squared:  0.8947,    Adjusted R-squared:  0.8791
F-statistic: 57.34 on 4 and 27 DF,  p-value: 8.335e-13
```

1. Net Yards *Allowed* per Attempt (NY/A)
2. Percentage of Drives Ending in Score (Sc%)
3. Takeaway Rate (TO%)
4. Yards per Play *Allowed* (Y/P)

Regression Results: Top 5 Teams in TeamScore

	Predicted Offense EXP	Predicted Defense EXP	TeamScore	Real-Life Context
#1 Seattle Seahawks	110.555	53.706	**164.261**	Statistically, most balanced team in NFL
#2 Indianapolis Colts	165.486	-18.481	**147.005**	Just lost their starting QB to a torn achilles
#3 Los Angeles Rams	108.394	34.129	**142.552**	Current Super Bowl betting favorite
#4 Detroit Lions	**133.269**	**-20.832**	**112.437**	**Lowest current record of these 5 teams**
#5 Green Bay Packers	109.747	-0.869	**108.878**	Acquired a new superstar at start of season, who may still be getting settled

Calculating TeamScore

```r
offense_raw$Pred_Off_EXPPG <- predict(offense_model)
```

```r
defense_raw$Pred_Def_EXPPG <- predict(defense_model)
```

```r
# Merge together by team name
library(dplyr)

combined <- offense_raw %>%
  select(Tm, Pred_Off_EXPPG) %>%
  left_join(defense_raw %>% select(Tm, Pred_Def_EXPPG), by = "Tm")

# Calculate TeamScore
combined$TeamScore <- combined$Pred_Off_EXPPG + combined$Pred_Def_EXPPG

# Standardize TeamScore
combined$TeamScore_z <- scale(combined$TeamScore)

# Rank by TeamScore
combined_ranked <- combined[order(-combined$TeamScore), ]

combined_ranked$Rank <- seq_len(nrow(combined_ranked))

head(combined_ranked[, c("Rank", "Tm", "Pred_Off_EXPPG", "Pred_Def_EXPPG", "TeamScore")], 32)
```

	Tm	Pred_Off_EXPPG	Pred_Def_EXPPG	TeamScore
	Seattle Seahawks	110.555150	53.7058922	164.261043
	Indianapolis Colts	165.486342	-18.4812548	147.005087
	Los Angeles Rams	108.393808	34.1286201	142.522428
	Detroit Lions	133.269330	-20.8324557	112.436875
	Green Bay Packers	109.746727	-0.8688991	108.877828
	Houston Texans	46.717820	54.2305480	100.948368
	New England Patriots	97.172769	-15.1893320	81.983437
	Denver Broncos	43.795520	24.2303057	68.025826
	Kansas City Chiefs	120.696943	-57.5648743	63.132069
	Buffalo Bills	102.974046	-42.0110710	60.962975
	Los Angeles Chargers	63.457450	-17.4597201	45.997730
	Chicago Bears	91.698979	-69.6969463	22.002033
	Atlanta Falcons	52.381634	-32.9145293	19.467104
	Jacksonville Jaguars	10.732164	6.6501549	17.382319
	San Francisco 49ers	55.576630	-40.0468385	15.529791
	Philadelphia Eagles	43.141674	-34.7937922	8.347882
	Dallas Cowboys	118.943300	-111.8742406	7.069059
	Pittsburgh Steelers	21.003381	-38.2198468	-17.216466
	Baltimore Ravens	49.820217	-67.4465875	-17.626370
	Tampa Bay Buccaneers	38.578143	-57.2412336	-18.663091
	Arizona Cardinals	9.169703	-53.7314116	-44.561709
	Miami Dolphins	21.470631	-79.4415952	-57.970964
	Cleveland Browns	-95.467969	25.4468598	-70.021109
	Carolina Panthers	-3.180367	-70.3966284	-73.576995
	Minnesota Vikings	-56.536296	-17.3156157	-73.851911
	New York Giants	37.776694	-124.2566750	-86.479981
	Washington Commanders	47.866203	-147.1268586	-99.260655
	New Orleans Saints	-63.562924	-36.1370608	-99.699984
	New York Jets	-26.773430	-81.9974840	-108.770914
	Las Vegas Raiders	-76.777301	-33.3301240	-110.107425
	Cincinnati Bengals	11.676123	-127.8459694	-116.169846
	Tennessee Titans	-101.523096	-90.4513363	-191.974432

Conclusions

1

Football is hard to predict

Tons of major external factors that contribute to team success

2

Correlation does not equal causation

Just because we found relationships between of our variables, doesn't necessarily mean they are causally related.

3

The human element

All of this data is about real people, who play a complex game that is constantly evolving

4

Direction > Definition

The best practice for interpretation of this data is to guide, enhance, and support operations, rather than redefine how football is analyzed

Thank you