Has the NFL Become a Passing League?

Andrew Lee

INF 6490: Statistics and Data Analysis

Professor Smith

December 12, 2025

Introduction

My dataset was titled *NFL Passing Statistics (2001-2023),* and I downloaded it from Kaggle. The dataset contains NFL passing statistics for every game since 2001, containing a record of every player who attempted a pass within that time. It was uploaded by Rishab Jadhav and represented the most comprehensive dataset of passing statistics that I could find.

The primary objective of my analysis is to answer the following question: has the NFL become a passing league? To answer this question, I decided to examine how many pass attempts there are each year, what the average yards per attempt has been each year, and what the average passer rating has been each year. I also wanted to know how passer rating is impacted by attempts and yards per attempt. To accomplish this, I needed to group the data by year and examine the variables of *ATT*, *Y/A,* and *Rate*.

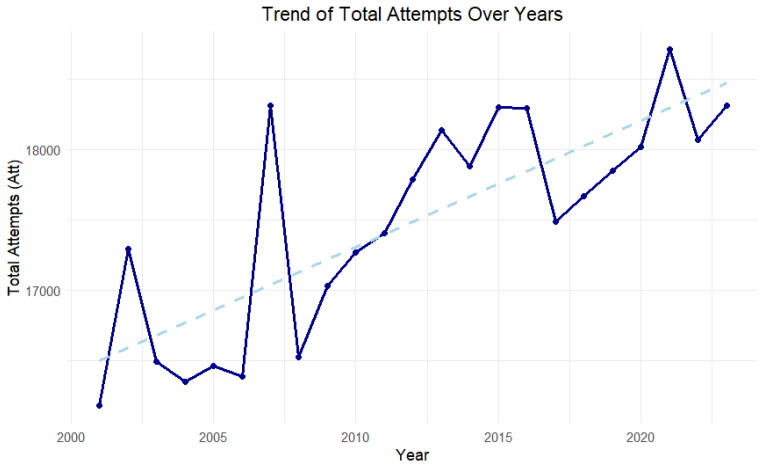
Process

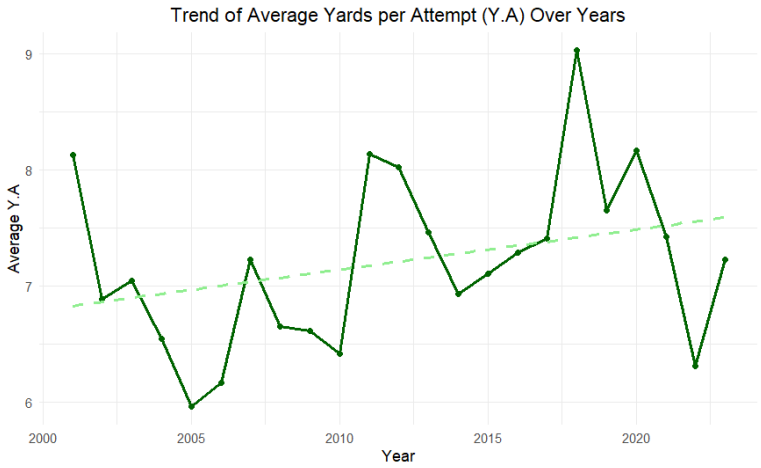
I started by loading all packages I would need and the CSV file for the dataset. I then created a summary of the data and checked for any missing values (none were found). I then ran a code to group the passing data by the categorical variable of *Year*. This allowed me to see the variables I was examining for each season from 2001-2023.

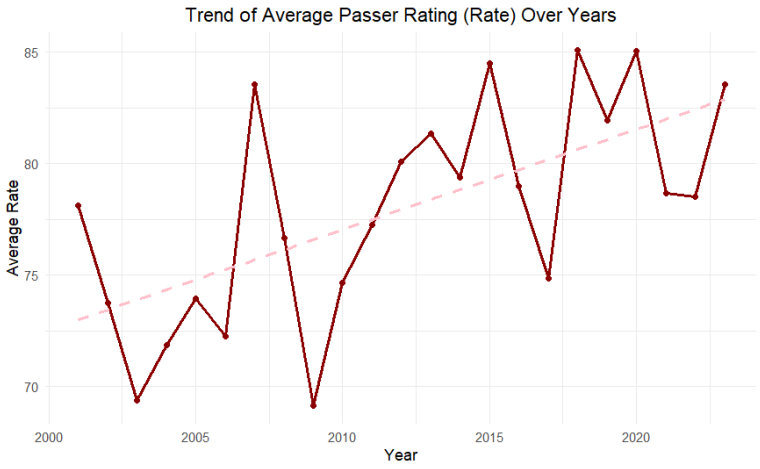
Next, I computed some descriptive statistics. The table below shows those statistics for the variables of *ATT, Y/A,* and *Rate*.

Variable	Mean	Median	Variance	SD
Att	171.16936	48.00	44229.7457	210.308691
Y.A	7.20400	6.60	55.7424	7.466083
Rate	77.79183	80.05	1020.7794	31.949639

I then created some data visualizations to see what these variables looked like year by year. I created a line graph for each variable by season with a trend line. Those visualizations are included below.







I then conducted tested the relationships using Pearson/Spearman correlation and ran Time Series Regression analyses to test trends in the data. I finished by calculating a 95% confidence interval for passer rating. The results are included below.

```
--- Pearson Correlation Matrix ---
             Att         Y.A      Rate
Att   1.00000000 -0.01791832 0.2118592
Y.A  -0.01791832  1.00000000 0.6662844
Rate  0.21185922  0.66628436 1.0000000

--- Spearman Correlation Matrix ---
           Att       Y.A      Rate
Att  1.0000000 0.3127667 0.2242439
Y.A  0.3127667 1.0000000 0.8274637
Rate 0.2242439 0.8274637 1.0000000


--- Time Series Regression: Avg Rate vs. Year (2001-2023) ---

Call:
lm(formula = Avg_Rate ~ Year, data = yearly_summary)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4486 -2.7354  0.3261  2.5514  7.8462

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -825.8917   244.7360  -3.375  0.00286 **
Year           0.4492     0.1216   3.693  0.00135 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.87 on 21 degrees of freedom
Multiple R-squared:  0.3937,      Adjusted R-squared:  0.3649
F-statistic: 13.64 on 1 and 21 DF,  p-value: 0.00135


--- Time Series Regression: Total Attempts vs. Year (2001-2023) ---
```

```
Call:
lm(formula = Total_Att ~ Year, data = yearly_summary)

Residuals:
   Min     1Q Median     3Q    Max
-604.9 -339.2 -187.4  358.0 1273.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -162684.36   30773.40  -5.287 3.06e-05 ***
Year             89.55      15.29   5.855 8.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 486.6 on 21 degrees of freedom
Multiple R-squared:  0.6201,      Adjusted R-squared:  0.602
F-statistic: 34.28 on 1 and 21 DF,  p-value: 8.22e-06

n--- Time Series Regression: Avg Yards per Attempt (Y.A) vs. Year
  (2001-2023) ---

Call:
lm(formula = Avg_YA ~ Year, data = yearly_summary)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24606 -0.40262 -0.05647  0.20869  1.61017

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -62.35721   46.02272  -1.355    0.190
Year          0.03458    0.02287   1.512    0.146

Residual standard error: 0.7277 on 21 degrees of freedom
Multiple R-squared:  0.09813,     Adjusted R-squared:  0.05518
F-statistic: 2.285 on 1 and 21 DF,  p-value: 0.1455



--- 95% Confidence Interval for Overall Mean Passer Rating (Rate) ---
Mean Rate (Overall): 77.79
95% CI: [ 76.5 ,  79.08 ]
```

I then trained and built a linear regression model and ran it on the test set. I have also

included these results below.

```
Training Data Size: 1689
Testing Data Size: 661


--- Linear Regression Model Summary ---

Call:
lm(formula = Rate ~ Att + Y.A, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-126.767   -9.581   -1.292   10.412   69.278

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.145437   0.913507   53.80   <2e-16 ***
Att          0.035154   0.002668   13.18   <2e-16 ***
```

```
Y.A          3.145148   0.079108   39.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.79 on 1686 degrees of freedom
Multiple R-squared:  0.5083,      Adjusted R-squared:  0.5077
F-statistic: 871.4 on 2 and 1686 DF,  p-value: < 2.2e-16


--- Model Performance on Test Set ---
Test R-squared: 0.4362
Root Mean Squared Error (RMSE): 22.93
```

Interpretation

When reviewing the data visualizations, the slope of the trend lines makes it clear that *ATT* and *Rate* have both increased at a statistically significant rate. This is confirmed by the P-values found through the Time Series Regression. The rate of increase for the variable of *Y/A* cannot be considered statistically significant. This tells us that there is not only much more passing volume over time in the NFL, but also that the passer performance has increased over that time as average passer rating has increased.

The correlation analysis reveals some other interesting relationships between the variables. Through the Pearson score, we can see that *ATT* has a weak relationship with *Rate*, meaning that more attempts do not drive higher passer ratings. This is in contrast to the relationship between *Y/A* and *Rate*, which shows a strong relationship between increased yards per attempt and increased passer rating.

The Linear Regression Model explains about 51% of the variance in passer rating, which is meaningful but tells us that about half of the variance is being controlled by other variables we have not accounted for. It also shows that both *ATT* and *Y/A* are statistically significant predictors, although the significance is much higher for *Y/A*. In fact, the model shows that yards per attempt is roughly 90 times more impactful on passer rating than attempts, as a 1

unit increase in *Y/A* results in approximately a 3.15 point increase in *Rate* compared to a 1 unit increase in *ATT* resulting in only a .04 point increase in *Rate*.

There are limitations to the model I created, primarily regarding missing variables that are key in determining passer rating. I believe including variables such as *TD, INT,* and *SACK* would have better explained the variance in passer rating.

Summary

Overall, the answer to the question about whether the NFL has become a passing league is a resounding yes. Volume and performance have steadily increased at a statistically significant level. Passer rating is not correlated with volume but is highly correlated with yards per attempt. This means that if NFL quarterbacks increase their yards per attempt and become more aggressive with their passing attempts in the future, they can expect to see a rise in their passer rating and better performance on the field.

Appendix A: Code for Final Project Analysis

```
---
title: "LeeFinalProject"
author: "Andrew Lee"
date: "2025-12-10"
output: html_document
---
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```{r}
library(ggplot2)
library(dplyr)
library(readr)
library(janitor)
library(ggthemes)
library(plotly)
library(caTools)
```
```{r}
passing_data <- read.csv("passing_cleaned.csv")
```
```{r}
head(passing_data)
summary(passing_data)
str(passing_data)
cat("\n--- Missing Values Check ---\n")
print(colSums(is.na(passing_data)))
```
```{r}
yearly_summary <- passing_data %>%
 group_by(Year) %>%
 summarise(
  Total_Att = sum(Att, na.rm = TRUE),
  Avg_YA = mean(Y.A, na.rm = TRUE),
  Avg_Rate = mean(Rate, na.rm = TRUE),
```
```

```
 ) %>%
 ungroup()
head(yearly_summary)
```

```{r}
continuous_vars <- c("Att", "Y.A", "Rate")
calculate_summary_stats <- function(data, column) {
 data %>%
  summarise(
   Mean = mean(.data[[column]], na.rm = TRUE),
   Median = median(.data[[column]], na.rm = TRUE),
   Variance = var(.data[[column]], na.rm = TRUE),
   SD = sd(.data[[column]], na.rm = TRUE),
   .by = NULL
  ) %>%
  mutate(Variable = column, .before = 1)
}
continuous_summary <- bind_rows(lapply(continuous_vars, calculate_summary_stats, data = passing_data))
print(continuous_summary)
```

```{r}
yearly_summary$Year <- as.numeric(yearly_summary$Year)
plot_att_year <- ggplot(yearly_summary, aes(x = Year, y = Total_Att)) +
 geom_line(color = "darkblue", linewidth = 1) +
 geom_point(color = "darkblue") +
 geom_smooth(method = "lm", se = FALSE, color = "lightblue", linetype = "dashed") +
 labs(
  title = "Trend of Total Attempts Over Years",
  x = "Year",
  y = "Total Attempts (Att)"
 ) +
 theme_minimal() +
 theme(plot.title = element_text(hjust = 0.5))
print(plot_att_year)
plot_ya_year <- ggplot(yearly_summary, aes(x = Year, y = Avg_YA)) +
 geom_line(color = "darkgreen", linewidth = 1) +
```

```
 geom_point(color = "darkgreen") +
 geom_smooth(method = "lm", se = FALSE, color = "lightgreen", linetype = "dashed") +
 labs(
  title = "Trend of Average Yards per Attempt (Y.A) Over Years",
  x = "Year",
  y = "Average Y.A"
 ) +
 theme_minimal() +
 theme(plot.title = element_text(hjust = 0.5))
print(plot_ya_year)
plot_rate_year <- ggplot(yearly_summary, aes(x = Year, y = Avg_Rate)) +
 geom_line(color = "darkred", linewidth = 1) +
 geom_point(color = "darkred") +
 geom_smooth(method = "lm", se = FALSE, color = "pink", linetype = "dashed") +
 labs(
  title = "Trend of Average Passer Rating (Rate) Over Years",
  x = "Year",
  y = "Average Rate"
 ) +
 theme_minimal() +
 theme(plot.title = element_text(hjust = 0.5))
print(plot_rate_year)
```

```{r}
cor_data <- passing_data %>%
 select(Att, Y.A, Rate)
cat("\n--- Pearson Correlation Matrix ---\n")
pearson_cor <- cor(cor_data, use = "pairwise.complete.obs", method = "pearson")
print(pearson_cor)
cat("\n--- Spearman Correlation Matrix ---\n")
spearman_cor <- cor(cor_data, use = "pairwise.complete.obs", method = "spearman")
print(spearman_cor)
```

```{r}
cat("\n--- Time Series Regression: Avg Rate vs. Year (2001-2023) ---\n")
```

```
trend_model_rate <- lm(Avg_Rate ~ Year, data = yearly_summary)
print(summary(trend_model_rate))
cat("\n--- Time Series Regression: Total Attempts vs. Year (2001-2023) ---\n")
trend_model_att <- lm(Total_Att ~ Year, data = yearly_summary)
print(summary(trend_model_att))
cat("n--- Time Series Regression: Avg Yards per Attempt (Y.A) vs. Year
 (2001-2023) ---\n")
trend_model_ya <- lm(Avg_YA ~ Year, data = yearly_summary)
print(summary(trend_model_ya))
```

```{r}
mean_rate <- mean(passing_data$Rate, na.rm = TRUE)
sd_rate <- sd(passing_data$Rate, na.rm = TRUE)
n_rate <- sum(!is.na(passing_data$Rate))
sem_rate <- sd_rate / sqrt(n_rate)
alpha <- 0.05
degrees_freedom <- n_rate - 1
t_score <- qt(1 - alpha/2, df = degrees_freedom)
lower_bound <- mean_rate - (t_score * sem_rate)
upper_bound <- mean_rate + (t_score * sem_rate)
cat("\n--- 95% Confidence Interval for Overall Mean Passer Rating (Rate) ---\n")
cat(paste("Mean Rate (Overall):", round(mean_rate, 2), "\n"))
cat(paste("95% CI: [", round(lower_bound, 2), ", ", round(upper_bound, 2), "]\n"))
```

```{r}
model_data <- passing_data %>%
 select(Rate, Att, Y.A) %>%
 na.omit()
set.seed(42)
split_index <- sample.split(model_data$Rate, SplitRatio = 0.70)
train_data <- subset(model_data, split_index == TRUE)
test_data <- subset(model_data, split_index == FALSE)
cat("Training Data Size:", nrow(train_data), "\n")
cat("Testing Data Size:", nrow(test_data), "\n")
```

```{r}
```

```
regression_model <- lm(Rate ~ Att + Y.A, data = train_data)
cat("\n--- Linear Regression Model Summary ---\n")
model_summary <- summary(regression_model)
print(model_summary)
```

```{r}
test_predictions <- predict(regression_model, newdata = test_data)
calculate_test_rsq <- function(actual, predicted) {
  sst <- sum((actual - mean(actual))^2)
  sse <- sum((actual - predicted)^2)
  rsq <- 1 - (sse / sst)
  return(rsq)
}
calculate_rmse <- function(actual, predicted) {
  rmse <- sqrt(mean((actual - predicted)^2))
  return(rmse)
}
test_r_squared <- calculate_test_rsq(test_data$Rate, test_predictions)
test_rmse <- calculate_rmse(test_data$Rate, test_predictions)
cat("\n--- Model Performance on Test Set ---\n")
cat(paste("Test R-squared:", round(test_r_squared, 4), "\n"))
cat(paste("Root Mean Squared Error (RMSE):", round(test_rmse, 2), "\n"))
```