Spring 2021

NFL Penalty Analysis, Referee Influence and Penalty Trends Over Time

Zachary McDaniel
University of Iowa

NFL PENALTY ANALYSIS, REFEREE INFLUENCE AND PENALTY TRENDS OVER TIME

by

Zachary McDaniel

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in the Business Analytics and Information Systems

_____

Kristina Bigsby
Thesis Mentor

Spring 2021

All requirements for graduation with Honors in the
Business Analytics and Information Systems have been completed.

_____

John P. Murry, Jr.
Business Analytics and Information Systems Honors Advisor

NFL Penalty Analysis, Referee Influence and Penalty Trends Over Time

by

Zachary McDaniel

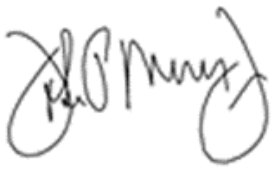
A thesis submitted in partial fulfillment of the requirements for graduation with Honors in the Tippie College of Business

Prof. Kristina Gavin Bigsby Thesis

Advisor

Spring 2021

Thesis received in partial satisfaction of the requirements to graduate with Honors in the Major



Prof. John Murry

Tippie College of Business Honors Director

NFL Penalty Analysis, Referee Influence and Penalty Trends Over Time

Zachary McDaniel

Henry B. Tippie College of Business

University of Iowa, Iowa City, Iowa 52242

Zachary-McDaniel@uiowa.edu

May 2021

Abstract

One of the biggest determinants of how American football is played are the rules which dictate what actions within the game are legal. Violations result in penalties of varying degrees, which can significantly impact the course of a game. This thesis examines NFL penalties over the last twenty years, focusing on understanding the effect of individual referees and home-field bias, changes in penalties over time, and the differences between NFL teams' penalties. While a statistical analysis did not find evidence supporting individual referee bias, there was a significant decrease in penalties for the years 2005 through 2008 . Additionally, there was a consistent and significant difference in penalties by team, with the Las Vegas Raiders, Baltimore Ravens, and Los Angeles Rams as the most-penalized teams, and the New York Jets and Indianapolis Colts as the least-penalized. Overall, the findings suggest structural explanations for penalty trends, such as major rule changes and organizational culture, rather than individual referee influence.

1 Introduction

The National Football League (NFL), established in 1920, is currently the most popular spectator sport in the United States. Recently, the NFL has distinguished itself among professional sports leagues for its ability to financially overcome the pandemic by relying on its near $10 billion in media contracts (Dixon, 2020). The prominent role of the NFL in American entertainment and society has led to the league continuously evolving over the last 100 years alongside the rest of the country. In addition to broader advances like the racial integration of the NFL, the league has seen many rule, safety, and organizational changes.

Elements that are today viewed as the status quo, such as player substitutions and penalties for roughing the passer were not originally part of professional football. Over time, the NFL has instituted new rules and penalties with the goal of making "the contests fairer, safer and more entertaining" (Evolution of NFL Rules, n.d). Penalty rules finalize at the organizational level of the NFL but begin forming through the actions of players, coaches, and owners. When players commit actions in game that are considered by the league and its officials to unsafe, inappropriate, or harm the integrity of the game, the NFL must analyze these actions and determine a future course of action for handling them, often in the form of penalties.

The purpose of this thesis is to investigate penalty changes in the last twenty years and how the teams within the NFL vary in penalty trends. The connection between an individual referee and penalty trends is investigated as a referee plays an integral role in the calling of penalties.

The paper continues with a background of the NFL and a review of previous research related to referees and penalty calls. Sections 3 describes the data and methods of analysis. Sections 4 and 5 discuss the results, conclusions, limitations, and the implications for future research.

2 Background

2.1 Referee Impact

Referees have discretion to make calls that often have a significant impact a game, and there are several studies which investigate the impartiality and accuracy of referee decision making. For example, Dohmen and Sauermann (2016) reviewed more than 60 previous studies of referee bias and found evidence of home team preference in decisions related to stoppage time, yellow and red cards, and penalty kicks in professional soccer. In addition to home bias, other research has focused on the effects of crowd size (Downward & Jones, 2007), previous team success (Erikstad & Johansen, 2020), team reputation (Jones, Paull & Erskine, 2002), and player body type (Van Quaquebeke & Geissner, 2010) on referee decisions in professional soccer. While most previous research on officiating has focused on soccer, Rodenberg & Lim (2009) examined the phenomenon of "payback calls" over 654 games the National Basketball League (NBA).

Of course, such work is limited by the ability to quantify bias or to differentiate between "good" and "bad" calls. Previous research on officiating accuracy has largely depended on using experts to evaluate referee decisions after then end of the game based on video (Mascarenhas, Collins & Mortimer, 2005).

Despite the popularity of American football, the research on referees and penalty calls is limited compared to other popular sports where there are fewer officials, and their individual actions have more identifiable outcomes (Snyder & Lopez, 2015). Understanding the impact of referees in the NFL is difficult due to the large officiating team and the number of differing roles on the field. An NFL officiating crew consists of seven roles including referee, umpire, head linesman,

field judge, line judge, side judge, and back judge (NFL Football Operations, n.d). Each title comes with different responsibilities for watching the field and the development of plays, although penalties can be assessed by any crew member. When a referee makes a call, their central tenets are consistency, accuracy, and fairness (Simmons, 2011). NFL referees use higher levels of subjectivity and discretion when making calls in the beginning and end of games when penalties have more significant impacts and referees feel more pressure to maintain the flow of the game (Snyder & Lopez, 2015). In 1999 the NFL first introduced the instant replay with a challenge system, which helps officials make accurate calls in high impact moments during games. Due to the rapid advancement in technology in the NFL, the replay system has become even more thorough and includes dedicated replay officials with the critical role of ensuring reviews are handled consistently, efficiently, and correctly (NFL Replay Officials, n.d). Thus, NFL penalties prove to be difficult to analyze over time due to the nature of referee crews and roles as well as the rapid advancement of game technology.

2.2 Penalties & Rule Changes

There are almost fifty different penalties calls in the NFL (nflpenalties, 2020) which can be broken down into several categories. Penalties can occur pre-snap (before a play begins), during play, and post play each with different violations and different officials judging. Additionally, penalties can be broken down into degrees of penalty yards starting at five yards and building to ten- and fifteen-yard penalties. It is possible for several penalties to be called in one play, and for penalties to be called on both the offense and defense. Generally, players want to avoid penalties but for opportunistic reasons still violate rules, which can set the tone for a game and

intimidate opponents (Snyder & Lopez, 2015). By intentionally violating rules, players take advantage of the referee's discretion and willingness to call a penalty to gain an advantage on the field.

In the last few decades, safety has been the primary driving force of rules, with over fifty rule changes since 2002 alone (NFL, 2019). The Competition Committee in the NFL consists of nine members and is one of the most visible and influence sources for changes in the game (Evolution of NFL Rules, n.d). This committee listens to team owners, coaches, player safety committee, team surveys, players union, and other medical experts when considering new rules. The NFL is constantly searching for areas to improve the game and benefit its constituents if it preserves the games integrity. With the advancement of technology and health sciences, the NFL has made a great effort to protect their players by introducing and changing rules. Some examples of major safety rule changes include protecting the passer, helmet to helmet contact, horse collar tackle, and protecting defenseless players. Each new rule change shifts the dynamic of the game as players adapt their actions to follow rules. The introduction of new rules and changes influences the trend of penalties over the last twenty years and is reflective of how the teams in the league play.

3 Methods

3.1 Data Collection & Preparation

All data for this study was collected from the website Pro Football Reference, which contained records on every football game, team, and official since the 1999 season(Pro Football Statistics and History, n.d).

Using R programming (R Core Team, 2020), I wrote a script which crawled several pages across the website to scrape information on every referee, game, team, and season from 1999 to 2019. Some data cleaning was required to properly format values and resolve multivalued attributes like weather which included temperature, wind, wind chill, and humidity. After cleaning the data, I merged relevant data sets and had two data sets with 3,000 to 40,000 rows. one table for referee stats and another for game and team stats. Descriptions of the features for both game and referee data are contained in *Table 1* and *Table 2* (see Appendix).

3.2 Referee Analysis

The goal of the referee analysis was to determine to what degree individual referees influence the outcome of penalties in a game or season. I derived several features to compare each referee's seasonal stats to averages across the entire league to determine trends. The key features created were home bias, harshness, and high penalty. Home bias is a binary feature that tracks if the winning percentage for home teams in games officiated by that referee for a given year was greater than the league average. Harshness is a binary feature that records whether the referee's average yards per penalty for a given year was greater than ten. Ten was chosen because the penalty calls with high discretion often incur fifteen-yard penalties so referees with an average high than ten are assumed to call more fifteen-yard penalties and the average yards per penalty in an NFL game is around seven to eight yards with an average of six to seven penalties per game (nflpenalties, 2020). High penalty is a binary feature that measures whether a referee's average penalties per game for a given year was greater than the average penalty calls per game over all referees in the league.

I conducted an exploratory analysis of the referee season statistics. This included creating a correlation matrix to check for strong linear relationships among the features and calculating AUC-ROC curves. The AUC-ROC curve is a performance measure using true and false positive rates which tells how well an individual feature or model is capable of distinguishing between classes (Narkhede, 2018). High penalty, which differentiates referees that tended to call more penalties or work on officiating crews in games with more penalties, was treated as the target class.

After this exploratory process, I narrowed my focus to investigate how a referee's tenure, home bias, and harshness were related to the binary high penalty measure. Finally, t-tests were completed to compare the means of the selected features for high penalty vs. low penalty referees. A t-test is a statistical test to compare differences in the means of two independent groups (Qualtrics, 2020). The null hypothesis for the t-test is that both means are equal, e.g., that there is no difference in average years of experience for referees categorized as high penalty vs. low penalty. The null hypothesis is that the means are not equal. These descriptive analyses were intended to understand the role an individual referee has in the outcome of a game and its penalties.

3.3 Time Analysis

This thesis specifically utilizes on data on NFL games and penalties from 1999 to 2019. The last twenty years are distinguished from the rest of NFL history by the organizational advancements in technology and the progression of rule changes defining penalties and supporting player

safety. The goal of my time analysis was to understand trends in penalties and penalty yards over this period.

I began by aggregating the game data by year and visualizing the total penalty calls and penalty yards over time. To determine if certain years had significantly higher or lower penalties, I ran an ANOVA test. An ANOVA is a statistical test to compare the differences between the means of three or more independent groups (Qualtrics, 2021). The null hypothesis for the ANOVA test is that all group means are equal, e.g., that there is no difference in average penalties by year. The alternative hypothesis is that at least one mean is different. I also tested the general assumptions of ANOVA, including normality, equal variance, and independence. The Bartlett, Kruskal-Wallis, and Shapiro test were used to test the variance, degree of difference, and normality of the total penalty calls and yards by year.

As a follow-up analysis, I conducted a post-hoc Tukey test to identify statistically significant years. The Tukey test compares all possible pairs of means to determine which specific years are different (Glen, 2021). Identifying the years which are statistically significant from 1999 through 2019 provides an accurate understanding when searching for external factors influencing the greater penalty trends over time. Knowing which factors influence penalty trends can provide insight for the NFL on how future changes might shift the landscape of penalty calls.

3.4 Team Analysis

In addition to examining NFL penalties from the perspectives of individual referees and time, I also investigated differences across NFL teams. The goal of the team analysis was to identify which of the thirty-two NFL teams stands out and investigate potential causes for their penalty

trends. Penalty data was aggregated by team, including average penalties and penalty yards, as well as differentiating penalties earned when the team was play home vs. away. Visualizations provided insight into average penalty trends across the league as well as initial indications of specific teams that appeared to deviate from the league norms. To further investigate the significance of these findings, I conducted ANOVA test to compare the means of all thirty-two teams. I followed up with a post-hoc Tukey test to specifically identify which team comparisons were statistically significant.

4 Results & Implications

4.1 Individual Referee Influence

Although individual referees may have some influence over penalty outcomes of a game, there is little evidence proving identifiable outcomes of their officiating. My investigation into the relationship between a referee's categorization as high penalty and features like home bias, harshness, and tenure were generally inconclusive.

In the exploratory analysis, no significant correlations above .75 were found between the independent and target variables, seen in *Table 3*.

Table 3

Correlation matrix results

	home penalties	home wp	total penalties	penalty yards	avg penalties	avg yards	avg yard per penalty	home bias	harsh	high penalty	tenure
home penalties	1.00	-0.09	-0.02	-0.02	-0.07	-0.06	0.03	-0.10	0.01	0.01	0.03

home wp	-0.09	1.00	0.09	0.09	0.02	0.04	0.07	0.71	-0.01	0.00	-0.01
total penalties	-0.02	0.09	1.00	0.99	0.39	0.29	-0.13	-0.02	-0.07	0.25	0.23
penalty yards	-0.02	0.09	0.99	1.00	0.40	0.34	-0.02	-0.02	-0.06	0.26	0.25
avg penalties	-0.07	0.02	0.39	0.40	1.00	0.92	0.06	-0.02	-0.07	0.64	0.04
avg yards	-0.06	0.04	0.29	0.34	0.92	1.00	0.44	0.00	-0.01	0.57	0.07
avg yard per penalty	0.03	0.07	-0.13	-0.02	0.06	0.44	1.00	0.04	0.22	0.02	0.07
home bias	-0.10	0.71	-0.02	-0.02	-0.02	0.00	0.04	1.00	0.00	0.02	0.01
harsh	0.01	-0.01	-0.07	-0.06	-0.07	-0.01	0.22	0.00	1.00	-0.03	-0.02
high penalty	0.01	0.00	0.25	0.26	0.64	0.57	0.02	0.02	-0.03	1.00	0.01
tenure	0.03	-0.01	0.23	0.25	0.04	0.07	0.07	0.01	-0.02	0.01	1.00

The results of the t-tests are displayed in *Table 4.* While there was a not a significant difference in home bias and tenure for referees categorized as high penalty, there was a significant result for harshness. Because these features are binary, the t-tests use the mean proportions of high penalty and low penalty referees compared to home bias, harshness, and tenure proportions. home bias and tenure classification have no significant relationship to high penalty, but harshness is significant at ranges from ten to fourteen average yards per penalty call. Referees whose average yards per penalty for a given year was higher than ten also tended to call aboveaverage numbers of penalties per game as compared to the league average for that year. The relationship between harshness and high penalty indicates that referees in charge of high discretion fifteen-yard penalty situations and make the call are more likely to be higher than the league average. This distinguishes the ability for referees responsible for high discretion calls to have an impact that stands out from the rest of the crew.

Table [1]

T-test results

Feature	t-value	p-value
Home bias	0.41	0.68
Harshness	3.9	9.80E-05
Tenure	-0.62	0.53

Looking at the results of the AUC-ROC curve to determine how well other features distinguished referees that were categorized as having a high penalty vs. low penalty season, I found that harshness, home bias, and tenure all had AUC values near 0.5 (*Table 5*). Thus, these features fail to differentiate between high penalty and low penalty referee seasons any better than random chance. The lack of evidence found in this analysis underscores the difficulty of analyzing individual NFL referee bias and influence in comparison to many other sports where the official's actions have more direct and identifiable outcomes on the game.

Table 5

AUC-ROC for Classifying High Penalty

Predictor	AUC
Harshness	0.509775501
Home bias	0.508688485
Tenure 0-5	0.504591926
Tenure 5-10	0.50201386
Tenure 10+	0.502578065

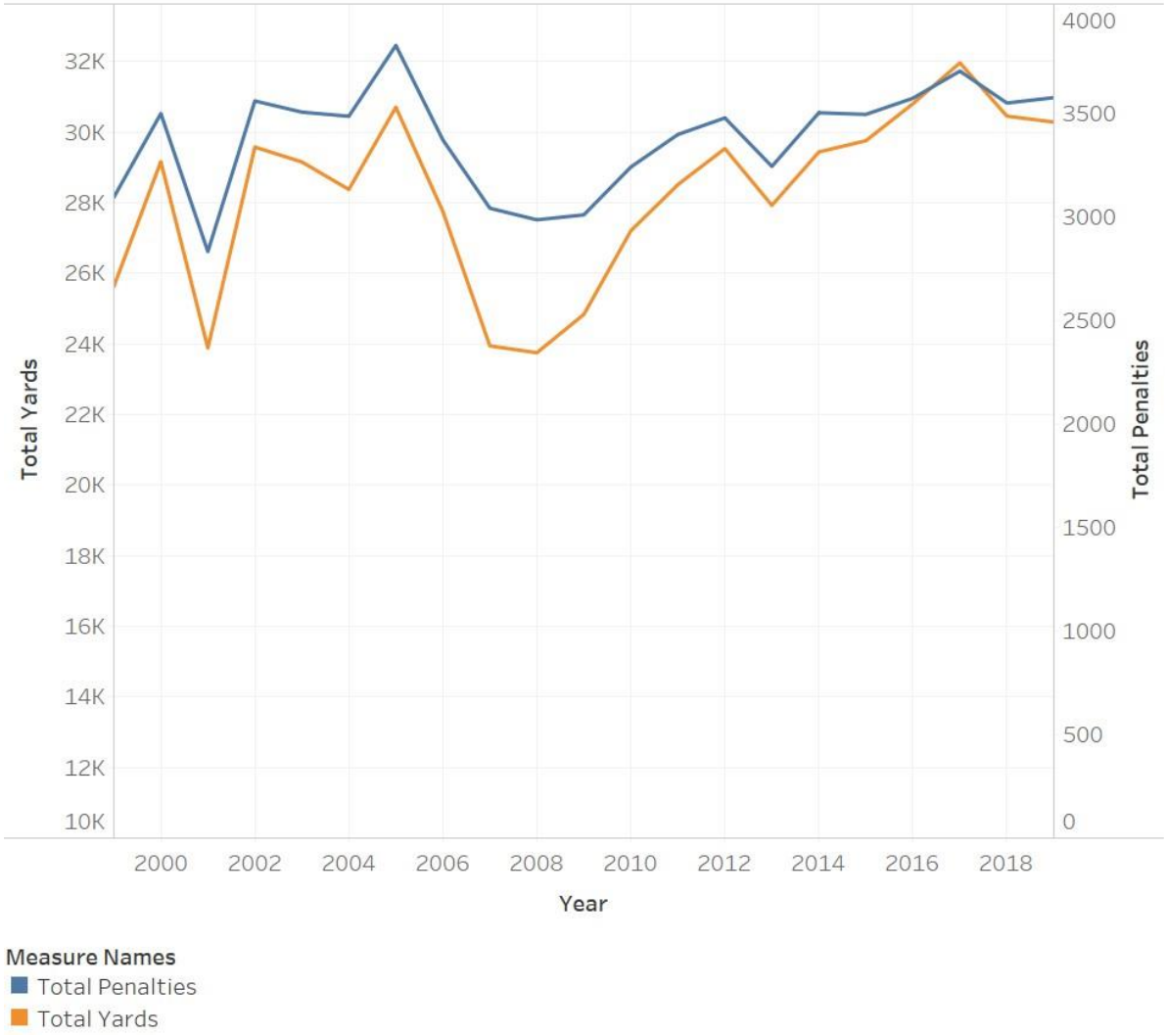
[1] .2 Significant Trends in Penalties

A visualization of the total penalties and total penalty yards assessed to all teams is displayed in *Figure 1.* While both overall trends are stable to slightly increasing over time, a few years stand

out. Specifically, the years 2001, 2007, 2008, and 2009 appear to have abnormally low numbers of penalty calls and yards. 2005 shows both the highest number of penalty calls and highest total penalty yards over all years in the data but over the next three years there is a 22% decline in penalties and yards. NFL rule changes are often reactionary to the game which could explain the peak in 2005 as a trend in dirty playing. After implementing new rules to offset the dirty or unsafe playing, a large drop off in penalties was observed in 2006, 2007, and 2008.

Figure 1

Total Penalties and Total Yards by Year (1999-2019)



An ANOVA test comparing the average total penalties by year resulted in a p-value of 2E-16, and the post-hoc Tukey test results are shown in *Table 6*. Although there were several years between 1999 and 2019 with significant differences, 2005 through 2008 were the most frequent and highest difference in average penalties per game.

Table 6

Tukey test output for average penalties per game by year

Comparison	diff	p adj
2007-2005	-2.83725	0
2008-2005	-3.04699	0
2015-2008	2.740543	2.43E-11
2008-2004	-2.70469	5.55E-11
2015-2007	2.530806	1.75E-09
2007-2004	-2.49495	3.42E-09
2005-2001	2.426766	1.82E-08
2006-2005	-2.23032	1.12E-07
2016-2008	2.2397	1.68E-07
2019-2008	2.209738	2.80E-07
2009-2005	-2.23872	2.89E-07
2018-2008	2.11236	1.41E-06
2010-2005	-2.09193	1.84E-06
2013-2005	-2.0807	2.21E-06
2015-2001	2.120319	5.03E-06
2016-2007	2.029963	5.24E-06
2008-1999	-2.09919	5.53E-06
2019-2007	2	8.35E-06

The biggest trends in the time 1999 to 2019 coincide with some of the largest rule changes occurring in the last twenty years. Many important rule changes occurred during this time regarding safety and many of these changes involved high discretion calls resulting in fifteenyard penalties. In 2005 the horse collar tackle was made illegal resulting in a fifteen-yard penalty and unnecessary roughness rules were expanded; in addition, 2006 and 2007 saw

expanded rules on horse collar, cut blocks, and other low hit scenarios (NFL Football Operations, n.d). In 2008, the incidental facemask ended with all facemask violations now resulting in fifteen-yard penalties (NFL Football Operations, n.d). This period of 2005 to 2008 was a pivotal time for rule changes in the NFL as they continued to progress their safety and health related rules, which has been an important factor in consideration of rules over the last twenty years.

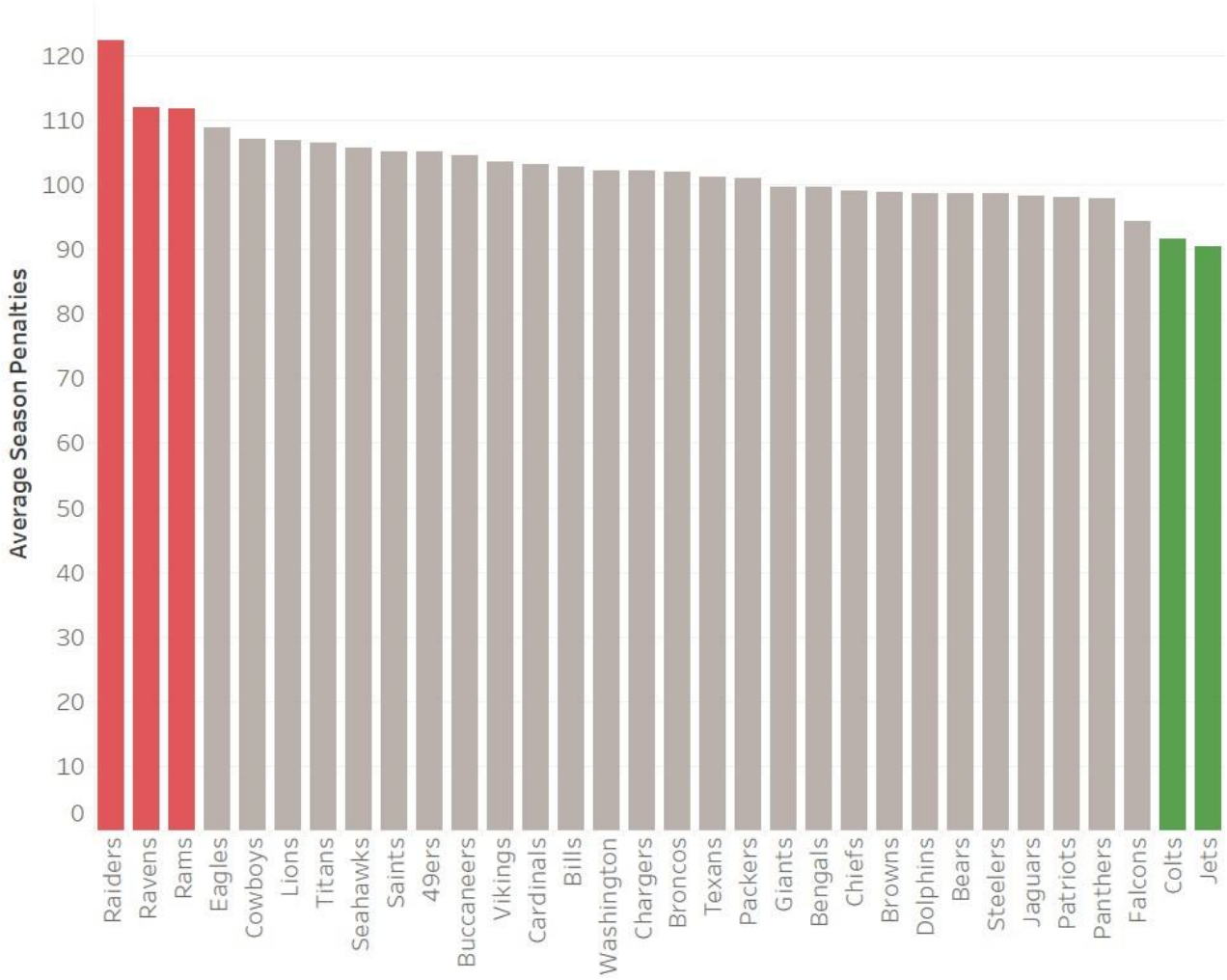
The sharp decreasing trend beginning in 2005 may be a result of these safety related rule changes. When a new rule is introduced like the outlaw of horse collar tackles, the decisions made at the organizational level can take time to become commonplace in the game. Additionally, as new rules define legal vs. illegal interactions on the field players must adjust their play styles to follow rules. The steep decrease in penalties observed from 2005 through 2008 is correlated to major rule changes occurring during the same period and should be considered by the NFL when making future rules that change how players can interact with one another and how to better integrate new changes at all levels.

4.3 Team Penalties

A visualization of average penalties per season for each of the 32 NFL teams, seen in *Figure 2*, provided insight into what the average team penalties were and which teams were the furthest from the league norm. Immediately, the following teams stood out as potential outliers: Las Vegas Raiders, Baltimore Ravens, Los Angeles Rams, Indianapolis Colts, and New York Jets. This result was unsurprising, as the Raiders have a reputation as one of the most notorious and most-penalized teams in the NFL (Arcellana, 2017).

Figure 2

Average season penalties by team



An ANOVA test of average penalties per season by team resulted in a p-value of 1.35E-07. The results of the post-hoc Tukey test indicated that the Raiders, Ravens, and Rams all had significantly higher average penalties in comparison to several other NFL teams. Specifically, the Raiders average penalties are significant when compared to almost two-thirds of the league (*Table 7*).

Table 7

Tukey test output for average penalties per season by team

Comparison	diff	p adj
New York Jets-Las Vegas Raiders	-33.9524	9.89E-08
Las Vegas Raiders-Indianapolis Colts	32.42857	5.77E-07
Las Vegas Raiders-Atlanta Falcons	29.57143	1.29E-05
Las Vegas Raiders-Carolina Panthers	26.04762	0.00041
Las Vegas Raiders-Jacksonville Jaguars	25.57143	0.000631
Las Vegas Raiders-Kansas City Chiefs	25.28571	0.000815
Las Vegas Raiders-Chicago Bears	25.2381	0.00085
New England Patriots-Las Vegas Raiders	-25.0476	0.001006
Las Vegas Raiders-Cleveland Browns	24.95238	0.001093
Miami Dolphins-Las Vegas Raiders	-24.5714	0.001521
Pittsburgh Steelers-Las Vegas Raiders	-24.5714	0.001521
Las Vegas Raiders-Cincinnati Bengals	24.09524	0.00228
New York Giants-Las Vegas Raiders	-23.5238	0.003658
New York Jets-Los Angeles Rams	-23.1429	0.004974
New York Jets-Baltimore Ravens	-22.7619	0.00672
Las Vegas Raiders-Green Bay Packers	22.7619	0.00672
Los Angeles Rams-Indianapolis Colts	21.61905	0.015917
Los Angeles Chargers-Las Vegas Raiders	-21.4762	0.01765
Las Vegas Raiders-Denver Broncos	21.33333	0.019552
Indianapolis Colts-Baltimore Ravens	-21.2381	0.020921
Las Vegas Raiders-Houston Texans	21.9127	0.023828
Washington Football Team-Las Vegas Raiders	-21	0.024726
Las Vegas Raiders-Buffalo Bills	21	0.024726
Las Vegas Raiders-Arizona Cardinals	20.61905	0.032109

The Raiders also were identified as the most penalized away team in the league with significant differences in average penalties by season in comparison to fifteen teams, or almost half the league, as seen in *Table 8.* Following the Raiders but with fewer significant comparisons, the Ravens and Rams also stood out for their higher-than-average total penalties and penalty yards in the last twenty years. At the bottom of the spectrum, the New York Jets were identified as the least penalized home team from 1999 to 2019 and the Indianapolis Colts are the least penalized away team for the same time.

Table 8

Tukey test output for average away penalties per season by team

Comparison	diff	p adj
Las Vegas Raiders-Indianapolis Colts	20.71429	9.59E-07
New York Jets-Las Vegas Raiders	-16.0476	0.001279
Las Vegas Raiders-Atlanta Falcons	15.52381	0.002544
Las Vegas Raiders-Jacksonville Jaguars	15.52381	0.002544
Pittsburgh Steelers-Las Vegas Raiders	-15.2857	0.003447
Miami Dolphins-Las Vegas Raiders	-15.0476	0.004643
New England Patriots-Las Vegas Raiders	-14.7619	0.006585
Las Vegas Raiders-Carolina Panthers	14.7619	0.006585
Las Vegas Raiders-Cincinnati Bengals	14.57143	0.008271
New York Giants-Las Vegas Raiders	-14.2381	0.01221
Las Vegas Raiders-Cleveland Browns	14.14286	0.013615
Minnesota Vikings-Las Vegas Raiders	-13.8095	0.01977
Indianapolis Colts-Baltimore Ravens	-13.4762	0.028327
Las Vegas Raiders-Kansas City Chiefs	13.42857	0.029786
Las Vegas Raiders-Chicago Bears	13.33333	0.032907
Las Vegas Raiders-Buffalo Bills	13.2381	0.036313

When comparing the Tukey test for home and away penalties, it is apparent that less teams stand out as significant for home penalties. The only team with significance were the Raiders, Ravens, and Rams which all were significant compared to the Jets (*Table 9*). The only consistent team appearing to cause these comparisons is the Raiders, whose abnormally high penalties triggered many of the findings for home and away penalties.

Table 9

Tukey test output for average home penalties per season by team

Comparison	diff	p adj
New York Jets-Las Vegas Raiders	-17.9048	1.42E-05
New York Jets-Los Angeles Rams	-17.1905	4.70E-05
Las Vegas Raiders-Atlanta Falcons	14.04762	0.005019
New York Jets-Baltimore Ravens	-13.9524	0.005685
Los Angeles Rams-Atlanta Falcons	13.33333	0.012431

It is notable that averaging each of the thirty-two team's penalties over twenty years still yields significant differences among the teams. Over a twenty-year span, a team goes through countless players, multiple coaches and owners, and some teams even change cities. There is consistency of penalties for teams like the Raiders and Jets on both ends of the penalty spectrum. This finding indicates persistent organizational culture encouraging certain kinds of play and a team's reputations sticks with them over time. If a team is perceived to be "bad", then referees may unconsciously pay more attention to that team or be more likely to interpret gray area plays in an unfavorable way (Jones, Paull & Erskine, 2002).

5 Conclusions & Future Work

This thesis explores penalties in the NFL from several perspectives. First, I examined the impact that referees have on penalty calls and the course of the game. Referee crews consist of seven types of officials, and these crew members can perform any role throughout the season. Thus, the dynamics of referee crews make quantifying the effect of individual referees challenging. I performed a descriptive analysis focused on several derived features comparing individual referees' performance by year with averages computed over the entire league. Ultimately, there were no conclusive results on the relationship between individual referees' average number of penalties per game, tenure, or home bias. However, there was initial evidence of a relationship between a referee being categorized as "high penalty" (having an average number of penalties per game above the league average) and "harsh" (average yards per penalty greater than ten). While it is difficult to distinguish an individual referee's impact as part of the crew, this result suggests that referees in charge of high discretion calls may have a greater ability to influence the outcome of the game.

Next, I investigated how penalties and penalty yards have changed over time and discovered an interesting trend from 2005 through 2008. The average penalties per game for these years found was significantly different when compared to other years in the data. Comparing this result to the evolution of NFL rules, it appears likely that substantial decrease in penalties and penalty yards for this period was related to major rule changes at that time, including the introduction of horse collar penalties and the end of incidental facemasks. Major rule changes that add new criteria on how players can interact with each other can impact penalty trends as seen in the peak during 2005 followed by a 22% decrease in penalties over the next three years. New rules are reactionary to the state of the game and implementation of the rules can lead to significant changes in league penalties. Even with constant rule changes, the overall trend in NFL penalties continues to rise. This positive trend could be explained by players and coach's ability to find new ways to engage and create grey area plays that have not yet had rules created for them. The NFL should consider a new process for how they implement new major rule changes to avoid trends like the 2005 through 2008 drop-off.

Finally, I analyzed the penalty data from at the team level, identifying which of the thirty-two teams in the NFL stood out for either their high or low penalty statistics. I found that Raiders were the most penalized team overall and in away games, which matches the public infamy given associated with the Raiders. The Baltimore Ravens and Los Angeles Rams were also significant as highly penalized teams over the last twenty years while the New York Jets and Indianapolis Colts were the least penalized teams in the NFL during this time. The findings of significant teams over this twenty-year period indicates that despite the ongoing changes teams experience, many teams have consistent penalty performance. This may be a result of a sticky

reputation and unconscious bias from referees against those teams or persistent organizational culture leading to a style of play. These findings indicate that individual teams do have the ability to influence penalty trends by their consistent seasonal penalty performance.

Overall, the findings suggest structural and organizational mechanisms behind penalty trends, such as the ongoing evolutions of rules and differences between the thirty-two NFL teams, rather than individual referee decisions and biases.

5.1 Limitations

In this analysis, there are a few limitations that can be improved upon in future research regarding NFL penalties. First, the lack of performance metrics for officials and penalty calls makes it difficult to quantify referees' accuracy or bias. Based on the data available, I was able to examine information on penalties and penalty yards for home and away teams at both the game and seasonal level but a new performance metric on call accuracy would strengthen arguments for how the harshness of an individual referee can impact the outcome of a game. Second, the referee-level analysis treated referees from 1999 to 2019 as one cohesive group, yet the analysis of penalties over time found significant variation in total penalties and penalty yards by year. Thus, finding significant results may be more difficult due to these trends, ongoing rule and organizational changes in the NFL, and referee role changes throughout games and seasons. Third, a limitation of the time analysis was the method used to define years. Specifically, games were assigned year labels based on calendar date, and not based on the NFL season. Because the NFL regular season starts in September and ends in January (with the postseason ending in February), this resulted in games occurring in the beginning of a new

calendar year being counted with games for the upcoming season. Fourth, the referee data collected for this project was measured at the season level, lacking data on individual penalties.

More granular data would be useful to further investigate the connection between rule changes and the penalty trends over time and differences by team. An analysis with additional information on penalty types would provide further insight into the direct impact of rule changes. Information on penalty type and what quarter the call was made would also strengthen my analysis of team penalties by understanding how teams are penalized and what trends have occurred across the league over time. Finally, the current data contains no information on players or coaches. While the team analysis identifies team penalties that are significant in comparison to their peers, it would benefit from further investigation into why those teams are significant, and whether trends in team penalty statistics can be related to organizational leadership.

5.2 Future Work

Although this thesis examines NFL penalties and referees from several perspectives, there is more work to be done in this field. Because football stands out among professional sports due to the size and composition of officiating crews, a deeper investigation into how the sevenperson referee crew directly influences the penalty outcomes of the game would be an excellent addition this domain. Additionally, more work is needed to understand the relationship between the introduction of new rules and penalty statistics. The development of metrics to define accuracy and bias of referee decisions, especially considering recent innovations in replay technology, would represent huge step forward in the research on penalties and officiating in the NFL.

Appendix

Table 1

Official Stats Data Dictionary

Official Stats contains 2982 rows with 20 variables. This data frame contains information on all officials in the NFL since 1999 and their penalty statistics per year compared the league average that year.

Feature	Type	Definition
URL	Char	Web address for official's information on Pro Football Reference
Officials	Char	The first and last name for official
Start year	Int	Year that official started working for the NFL
End year	Int	Year that official ended working for the NFL
Years	Int	Year that statistics occurred in
Position 1	Char	Main role for official in that year
Position 2	Char	Secondary role for official in that year
Playoff games	Int	Count of playoff games officiated in that year
Total penalties	Int	Count of penalties called by official in that year
Avg penalties	Int	Average penalties called by the official per game in that year
League avg penalties	Int	Average penalties per game in that year over entire league
Home	Int	Count of penalties called on home team in that year
Away	Int	Count of penalties called on away team in that year
Home penalties	Num	Percentage of penalties called on home team in that year
League home penalties	Num	Average percentage of penalties called on home team in that year over entire league
Penalty yards	Int	Number of yards resulting from penalties called by official in that that year
Avg yards	Num	Average yards resulting from penalties called by official in that year

League avg yards	Num	Average penalty yards per game in that year over entire league
Home win %	Num	Percentage of games officiated in that year won by home team
League home win %	Num	Average percentage of games in that year won by home team over entire league

Table 2

Combined Game Info Data Dictionary

Combined Game Info contains 40509 rows with 22 variables. This data frame contains information on every single NFL game since 1999 including penalty statistics and important information about the conditions during the game. Each game has information for every referee involved.

Feature	**Type**	**Definition**
Game URL	Char	Web address for game's information on Pro Football Reference
Date	Date	Game date
Won toss	Char	Team that won the beginning toss
OT toss	Char	Team that won the overtime toss (if applicable)
Roof	Factor	Type of roof in stadium where game was played
Surface	Factor	Type of field where game was played
Duration (min)	Int	Game duration (minutes)
Attendance	Int	Game attendance
Temp	Int	Temperature during game (Fahrenheit)
Humidity	Int	Relative humidity during game
Wind	Int	Wind speed during game (mph)
Wind chill	Int	Wind chill during game (Fahrenheit)
Position	Char	The position each official played during the game
Home team	Char	Home team
Home points	Int	Points scored by the home team
Home penalties	Int	Count of penalties called on home team
Home yards	Int	Number of yards resulting from penalties called on home team

Away team	Char	Away team
Away points	Int	Points scored by away team
Away penalties	Int	Count of penalties called on away team
Away yards	Int	Number of yards resulting from penalties called on away team

References

Code Repository: https://github.com/ZachMcDaniel/NFL_penalty_analysis

ANOVA (Analysis of Variance): Definition & Methods // Qualtrics. Qualtrics. (2020, September 22). https://www.qualtrics.com/experience-management/research/anova/

Arcellana, J. (2017, October 3). *Oakland Raiders Are The Most Penalized Team In The NFL, Need Discipline to Win*. Bleacher Report. https://bleacherreport.com/articles/474386-oaklandraiders-have-been-plagued-with-penalties-can-not-win-without-discipline

Dixon, S. (2020, June 13). *TV money gives NFL leg up if fans can't fill teams' coffers*. AP NEWS. https://apnews.com/article/virus-outbreak-nhl-sports-general-green-bay-packers-mlbc3b89b134160748ad616096bc3cc9166

Downward, P & Jones, M. (2007). Effects of crowd size on referee decisions: Analysis of the FA Cup. *Journal of Sports Sciences, 27 (14).* https://doi.org/10.1080/02640410701275193

Erikstad, M.K. & Johansen, B.T. (2020). Referee bias in professional football: Favoritism toward successful teams in potential penalty situations. *Frontiers in Sports and Active Living, 27.* https://doi.org/10.3389/fspor.2020.00019

Evolution of the NFL Rules: NFL Football Operations. Evolution of the NFL Rules | NFL Football Operations. (n.d.). https://operations.nfl.com/the-rules/evolution-of-the-nfl-rules/.

Glen, S. (2021, January 1). *Tukey Test / Tukey Procedure / Honest Significant Difference*. Statistics How To. https://www.statisticshowto.com/tukey-test-honest-significantdifference/.

Health & Safety Rules Changes. NFL Football Operations. (n.d.). https://operations.nfl.com/therules/rules-changes/health-safety-rules-changes/.

Introduction to T-test Theory and Use Cases // Qualtrics. Qualtrics. (2020, September 22). https://www.qualtrics.com/experience-management/research/t-test-analysis/.

Jones, M., Paull, G. & Erskine, J. (2002). The impact of a team's aggressive reputation on the decisions of association football referees. *Journal of Sports Sciences 20:* 991-1000. https://doi.org/10.1080/026404102321011751

Mascarenhas, D.R.D., Collins, D., & Mortimer, P. (2005). The accuracy, agreement and coherence of decision-making in rugby union officials. *Journal of Sport Behavior 28(3):* 253-271.

Narkhede, S. (2021, January 14). *Understanding AUC - ROC Curve*. Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

NFL. (2019, September 26). *NFL Health and Safety Related Rules Changes Since 2002*. NFL.com. https://www.nfl.com/playerhealthandsafety/equipment-and-innovation/ruleschanges/nfl-health-and-safety-related-rules-changes-since-2002

NFL Officials' Roles and Responsibilities. NFL Football Operations. (n.d.). https://operations.nfl.com/officiating/the-officials/officials-responsibilities-positions/.

NFL Penalties - 2020 League Penalty Stats - View by Total. NFL Penalty Stats Tracker. (n.d.). https://www.nflpenalties.com/

NFL replay Officials. (n.d.). https://operations.nfl.com/officiating/instant-replay/nfl-replayofficials/.

Pro Football Statistics and History. Pro. (n.d.). https://www.pro-football-reference.com/.

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/

Rodenberg, R.M. & Lim, C.H. (2009). Payback calls: A starting point for measuring basketball referee bias and impact on team performance. *European Sports Management Quarterly 4*. https://doi.org/10.1080/16184740903331853

Simmons, P. (2011). Competent, dependable and respectful: Football refereeing as a model for communicating fairness. *Ethical Space: The International Journal of Communication Ethics, 8(3), 33-42.* *https://researchoutput.csu.edu.au/ws/portalfiles/portal/8800540/PID28009manuscript.pdf*

Snyder, K. & Lopez, M. (2015). Consistency, accuracy, and fairness: a study of discretionary penalties in the NFL. *Journal of Quantitative Analysis in Sports*, *11*(4), 219-230. https://doi-org.proxy.lib.uiowa.edu/10.1515/jqas-2015-0039

Stephanie. (2021, January 1). *Tukey Test / Tukey Procedure / Honest Significant Difference*. Statistics How To. https://www.statisticshowto.com/tukey-test-honest-significantdifference/.

Van Quaquebeke, N. & Geissner, S.R. (2010) How embodied cognitions affect judgments: Height-related attribution bias in football foul calls. *Journal of Sport and Exercise Psychology 32(1):* 3-23. https://doi.org/10.1123/jsep.32.1.3