Final Year Project

Analytics and the National Football League (NFL)

Michael Mullin

Student ID: 17360573

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Barry Smyth

UCD School of Computer Science
University College Dublin

April 30, 2022

Table of Contents

Terminology

In order to assist and support the reader's understanding of this research, terminology which is specific to the sport of American Football is explained hereunder.

- **Down**: A down is one of the most important concepts in the game of American Football. A down is another name for a play. On each down a play is made. The team in possession has four downs to advance ten or more yards towards their opponent's goal line. Each down situation is denoted by the down count plus the required yardage to gain. e.g. 1st-10. On each down the team in possession decides to run or pass. If the team gains the required yardage within 4 downs, the down counter resets back to one.
- **4th Down**: The 4th Down is the last attempt the team in possession has to gain the required yardage towards their opponent's goal line. 4th Downs are a critical aspect of the game and the decision a team makes on 4th Down has a strong influence on the outcome of the game. If the team in possession fails to gain the required number of yards on 4th Down, the opposition team receives possession from that position. A successful 4th down is defined as gaining the required number of yards. The three options teams can choose between on 4th Down are to "go for it" (run or pass), "punt" or try a "field goal" which are explained in detail in section 4.2.

A further in-depth explanation of all related terminology and associated information in this project can be found in section 8.2.

Chapter 1: **Introduction**

Data analytics in the domain of professional sports is evolving every year. Sports teams have embraced technology and data in order to boost the performance of their teams. Coaches and players have access to vast amounts of training data and this information is increasingly being used as the basis for more considered data-driven coaching decisions. To the casual viewer, the players on the field are the only ones influencing the game and making decisions. However, new systems have been developed such as live GPS tracking and performance monitors which help translate what's happening on the field to measurable metrics. The introduction of the Video Assistant Referee (VAR) in major soccer leagues around the world has also been groundbreaking in correcting on-field decisions made by referees. However, despite this growth in technology, 'old school' coaching is still present amongst the more traditional head coaches and team coordinators, particularly in the National Football League (NFL). The concept of home field advantage is a widely discussed phenomenon in all sports. While the existence of the home field advantage is not in dispute, the magnitude of the effect on a team and the precise causes are not completely clear. Additionally, in the past two years, COVID-19 has also posed many new challenges to both the individual players and their respective teams across the world.
This research project endeavours to investigate and explore the value of home field advantage to a team, while also studying and creating a predictive model for on-field decision making in the 4th Down, and the general impact of COVID-19 on the NFL is also explored.

A selection of some of the research in the area of home field advantage, 4th Down, and the impact of Covid 19 are briefly explored and outlined in Chapter 2. The project work plan and timeline are illustrated in Chapter 3. The three main research questions are introduced and elaborated in detail in Chapter 4. The methodologies and analytical frameworks pertinent to each of the research questions are explained in Chapter 5. Chapter 6 examines the results of the respective analyses and a discussion of the relevant findings. The concluding chapter summarises the main findings, limitations, and scope for future work and research in this area. Many of the findings that emerged in this small-scale project reflect the findings from other studies in this area.

Chapter 2: **Related Work**

Analytics in sports has come a long way in recent years with data being made more available to the public and people being encouraged to find patterns and useful information within the data. Many experts and fans have conducted their own individual analysis of various different aspects which have shown significant relevance to this project. By engaging with different articles and publications across various sports, the opportunity to acquire new knowledge and delineate gaps in research arose. Analysis of the National Football League (NFL) has become extremely prevalent within the community recently. Automatic Twitter accounts and live analytics such as NFL Next Gen Stats (Powered by AWS) have become the norm in what is a rapidly expanding area. The following related works were of particular interest as they provided insightful analysis and key findings into this field of research.

Home Field Advantage in Athletics: A Meta-Analysis [1]
Jamieson examined home-field advantage (HFA) by quantifying the probability of a home victory across all major sports. Overall, a significant advantage for home teams was observed across all sports (win probability = 0.6). Baseball represented the lowest win probability (0.56) while soccer showed the highest win probability (0.67) for home teams. This analysis is limited to examining each sport as a whole average. However, by analysing individual teams within sports, this could provide a clearer picture of how HFA deviates among the teams.

Home Field Advantage Analysis in the NFL [2]
Cadena breaks down the hypothesis that playing at home is an advantage in the NFL. He proposes that the way to get the real value of home field advantage is by analysing several factors such as strength of team and opponent, competence of the head coach and opposing head coach, quarterback and opposing quarterback performance, and how each team performs home v away. By investigating these factors, Cadena displays a significant difference in estimated home field advantage (eHFA) for each individual team. The Las Vegas Raiders (+2.4 eHFA) and Arizona Cardinals (+3.3 eHFA) represent the lowest and highest eHFA respectively. Cadena also shows how HFA has been declining in recent years (2006-2020). His work presents a very concise and informative analysis on HFA in the NFL. However, one of the main limitations to this study is the sample size is quite small, covering only the last 15 years.

NFL 4th Down Analysis [3]
Burke, a sports fan and math enthusiast compiled years of NFL game data to analyse what decisions teams should make on 4th Down. His analysis provides a basic strategy on what decisions to make on 4th Down. Burke uses expected points, net punt distance, and field goal percentage values to create a decision tree on what to do on 4th Down. A limitation of this study is the lack of analysis on key variables associated with 4th Down decision making such as game time remaining, score differential, and quarterback performance. Considering each of the factors could reach some alternate conclusions.

NFL 4th Down Model [4]
This Twitter account tweets every time a 4th Down occurs in the NFL. Each tweet analyses whether a team achieved the most optimal decision. The model utilises several variables including the teams playing, current score, game time remaining, distance to the end zone, and timeouts remaining. The model indicates the teams win percentages for each 4th Down outcome ("go for it", punt, kick a field goal). The model outputs a win percentage value for the recommended decision and a win percentage value for the actual decision the team makes. This model is very insightful and provides a clear picture of which teams are making the most optimal 4th Down decisions.

NFL Player Injury Rate During COVID-19 [5]
Due to the rising number of positive COVID-19 tests observed in NFL players during the summer of 2020, the NFL suspended the 2020 pre-season on July 21, 2020. This paper investigates the injury rate in NFL athletes during the first 4 weeks of the 2020 NFL season. There was a significant increase in the injury rate during weeks 1-4 of the 2020-2021 regular season for all comparisons with the injury rate both during the pre seasons and the regular seasons of three recent past NFL seasons. Given the unprecedented suspension of pre-season training and teams' inability to continue with practice and games due to social-distancing precautions, athletes were forced to

train and condition on their own. This paper speculates that in the absence of organized team activities, athletes developed sports de-conditioning leading to inevitable injury.

Referees in Professional Soccer During the COVID-19 Pandemic [6]
This article analyses the impact of COVID-19 on professional soccer games. Leitner and Richlan analyse referee performance and the variance of home field advantage during the pandemic. The analysis of yellow cards for fouls indicate that referees give less preferential treatment to the home team in ghost games than in regular games. The term "ghost game" refers to a game without fans in attendance. Secondly, the analysis shows that the absence of spectators significantly reduces the value of home field advantage. Leitner and Richlan show that due to the missing supporters in ghost games, referees perceived less social pressure from the home crowd, leading to minimal home advantage effect.

Performance Analysis in Rugby Union: A Critical Systematic Review [7]
Colomer et al. review the current state of performance analysis research in professional rugby union and consider the utility of common methods of analysing performance and the applicability of these methods within professional coaching practice. The paper highlights how performance analytics is lacking in the rugby union community. The majority of studies assessed lacked contextual information such as the opposition, match location, period within match, and field location. Colomer et al. illustrate how there is scope for the rugby union analysis research to improve by applying methods similar to what are in place in sports such as football and basketball.

Modelling the success rate of top English Premier League Football Clubs [8]
Oberstone describes a multiple linear regression model he developed that accounts for the relative success of English Premier League football clubs during the 2007-2008 season. The model is used to define six statistically significant team pitch factors that are key to a club's ultimate success as measured by points earned. The six key pitch actions defined are defending, goal attempts, discipline, passing, and crossing. This article deals with a large amount of data obtained from Opta Sportsdata, PA Sport-Actim Index, and Prozone.

Analysis of home field advantage during the Six Nations Rugby Championship [9]
This study investigates the effect of alternating home and away field advantage on selected performance indicators during the Six Nations Rugby Championship from 2005 to 2009. Vaz et al. sampled 75 games played over these five seasons to see whether teams obtained more favourable results when playing at home. Some of the factors taken into account in this study were final match result, number of tries scored, and number of passes. The results of the study show that teams score 50% more points when playing at home and in general, teams obtain more favourable results when playing at home.

Reading the above works helped provide an illustration of where the sports analytics community is at in terms of areas researched and aspects that have yet to be considered. Home Field Advantage is a topic that is relevant across all team sports and has been analysed in great detail. Jamieson has shown that HFA has a significant impact in the NFL but overall it varies from crucial to inessential across different sports. 4th Down analysis has only been looked at briefly as the NFL has entered the "modern area" of analytical-based decisions. Furthermore, by researching the sports analysis space, there is arguably scope for improvements to be made to the vast number of studies conducted to date.

Chapter 3: Project Work plan

3.1 Timeline

Project management commenced in early October with the allocation of project titles. The various tasks involved in the project were broken down into preliminary, writing, and computational sub-sections. While different sets of tasks will be worked on simultaneously, this division of work helps provide a clearer picture of the layout of the project. Figure 3.1 below illustrates the breakdown of the project load over the working period.

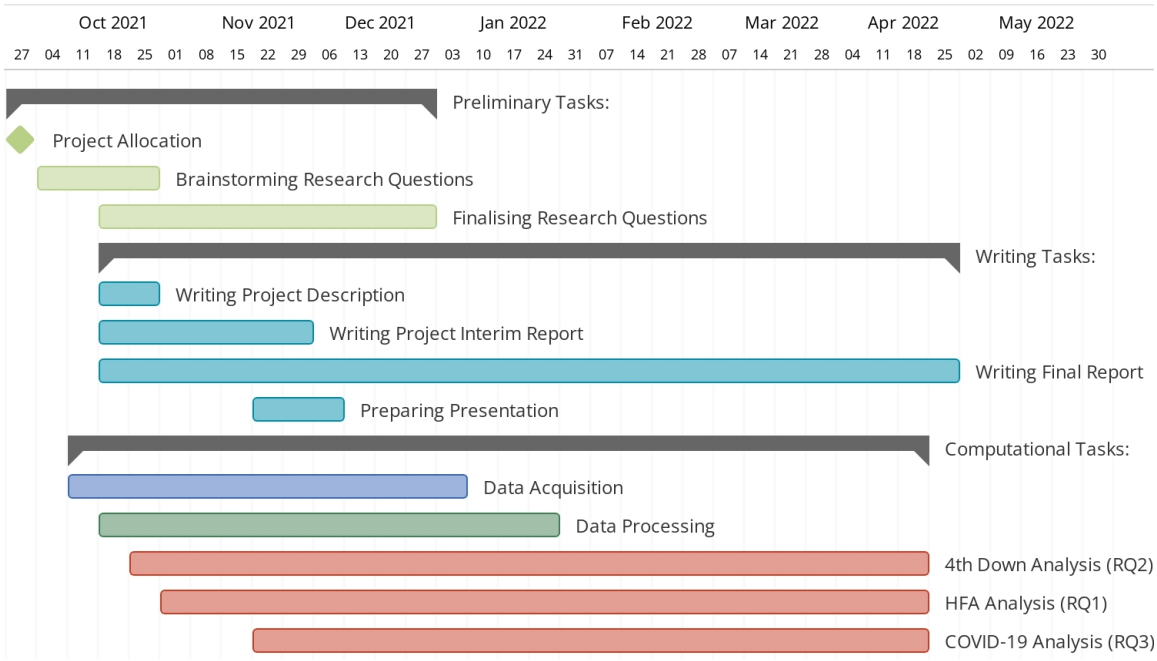


Figure 3.1: Gantt Chart

Preliminary Tasks
The initial brainstorming of possible research questions commenced once the project titles were allocated in early October. Finalising the research questions will be completed by early January.

Writing Tasks
A short two-page initial description of the project is due in late October. Six weeks has been allocated for writing the interim report which is due in early December. A presentation of the project is scheduled for Zoom delivery also in early December, two weeks has been allocated for the creation and rehearsal of this presentation. The final report is due on April 29th which is an extension of the interim report.

Computational Tasks
The appropriate data needed to carry out this project will need to be acquired. This entails researching and downloading appropriate datasets from various websites/sources. Once the data has been acquired it will need to be cleaned and processed before any analysis begins. Each of the research questions will be worked on simultaneously.

3.2 Data

There are two main core datasets used in this project. The Play by Play (PbP) dataset and the historical scores dataset. Several sub-datasets were created from these original datasets.

1. Play by Play Data

The PbP dataset was downloaded from the nflfastPy Github repository [10]. This dataset represents information for every play from every game from the 1999 season to the most recent 2021 season. The nflfastPy library was imported and each individual year dataset was loaded into a pandas dataFrame, as shown in the code below for the 2021 season.

```
import nflfastpy as nfl
df_2021 = nfl.load_pbp_data(2021)
```

Each individual year dataset, through a set of defined functions, was then concatenated together to form a master dataset for PbP data from 1999-2021. The master dataset consists of 373 columns and 1,098,570 rows. The sheer amount and detail of information present in the dataset was far greater than the amount needed for our analysis. A number of cleaning methods were carried out to ensure the data was complete, error-free, and easy to work with. The 4th Down sub-dataset was created from the master dataset by filtering for the value '4' in the 'down' column. For the purpose of analysing 4th Downs in the NFL it's important to recognize which 4th Downs are relevant for the approach. The term *Garbage Time* is a known term within in the NFL community and it refers to the final moments or minutes of a game in which one side has an insurmountable lead, substitutes often enter the game in place of starting players, and scoring is typically easier because of looser defensive play [11]. 4th Downs occurring within this period can skew the data whereby teams are making more aggressive decisions knowing that the game is already over. The *garbage time* for this dataset has been defined as when the win probability for a team is greater than 95% or less than 5%. Occurrences that meet this criterion were filtered out for this analysis.

```
df_wog = df[(df['wp'].between(0.05, 0.95)]
```

The penalties sub-dataset was also created from the master dataset by filtering the 'play_type' column for the string 'PENALTY'. A sample of some of the relevant columns within the master dataset are listed below:

home_team : Team playing at home. e.g. ATL

away_team : Team playing away from home. e.g. PHI

total_home_score : Home Team score.

total_away_score : Away Team score.

score_differential : Difference in scores from the team in possession's view.

yardline_100 : Distance to the opponents end zone (yards).

yards_gained : Yards gained on the play.

ydstogo : Yards required to gain a new set of downs.

year : The current year.

game_seconds_remaining : Seconds remaining in the game.

down : Current down count (1-4).

desc : Description of the previous play. e.g. (3:27) 12-T.Brady pass incomplete short left to 11-J.Edelman.

play_type : Type of play (pass, run, field goal or punt).

2. Historical Scores

The historical scores dataset was downloaded from Toby Crabtree's Kaggle repository [12]. This dataset represents information for every game dating back to the 1999 season. The dataset consists of 17 columns and 6,092 rows. The key information held in this dataset is historical team scores and the sportsbook lines. A sportsbook is a gambling establishment that takes wagers on sporting events and pays out winnings. Sportsbooks set informative lines for NFL games which define the expected total points (over/under line) and the expected margin that the favourite team will win by. Figure 8.5 shows an example of the lines set by a sportsbook. Each row in the dataset refers to a game played in the NFL. It is important to note that the home team for the Super Bowl (the final) alternates between teams each year. Interestingly, the Los Angeles Rams played the 2021 Super Bowl in their home stadium but were not deemed the home team. The original dataset obtained defined the Super Bowl home team correctly as per the rule mentioned above. However for the purpose of analysing home field advantage, the Super Bowl games were removed from the data. A sample of some of the relevant columns within the dataset are listed below:

team_home : Team playing at home. e.g. ATL

team_away : Team playing away from home. e.g. PHI

score_home : Home Team score.

score_away : Away Team score.

team_favorite_id : Favourite team defined by sportbooks.

spread_favorite : Margin (points) that the favourite is expected to win by defined by sportbooks.

over_under_line : Expected number of points in the game defined by sportbooks.

schedule_season : Current year.

schedule_week : Current week (1-18, Wildcard, Division, Conference, Super Bowl)

stadium_neutral : True/False on whether the game was played in a neutral stadium.

Despite the historical_scores dataset being substantially smaller in size relative to the PbP dataset, there was more cleaning and computational work required. Nineteen additional columns were created such as favourite score, underdog score, favourite win percentage, and whether the home team was favourite. The favourite win percentage was calculated by mapping the favourite margin to a win probability (Figure 8.4).

Chapter 4: Research Questions Overview

In this section we will introduce in detail our three main research questions.

4.1 RQ1: How Valuable is Home Field Advantage?

Home Field Advantage (HFA) can simply be viewed as any advantage a team or individual gains when playing in their home venue. The effects of home field advantage can be seen across various individual and team sports such as basketball [13], soccer [1], and rugby union [9]. Home field advantage in the National Football League (NFL) varies for each individual team. It could be argued that the Seattle Seahawks enjoy a bigger home-field edge than most teams because their stadium is considered one of the loudest in the league. In 2013, CenturyLink Field (Seattle stadium) broke the Guinness Book of World Records for the loudest crowd noise in their home game vs New Orleans Saints when they reached a deafening 137.6 decibels [14]. For context, 130dB can be attributed to the sound of a Jet taking off. Similarly, the Denver Broncos play at an altitude of 5,280 feet above sea level. The Broncos players live and train there, so they're used to this high altitude, but visiting teams have to adapt. The adaptation process supposedly takes one day for every 1,000 feet over 3,000 feet. The sports gambling industry and their enthusiasts are continually trying to discover the correct value of home field advantage in terms of "point-spread" to create gambling odds. The point-spread is a number created by oddsmakers at sportsbooks that serves as a handicap between two opponents. Panayotovich [15] explains that if two teams possessed similar ratings and were essentially mathematical equals on a neutral field, a sportsbook would round up and make the home team a 3-point favourite in the point-spread. This 3-point varies across different home teams per season and can be attributed to a number of factors that influence the strength of home field advantage:

- **Crowds**: Home crowds are one of the main factors contributing to home field advantage. The number, density, and behaviour of spectators can have a huge impact on the performance of teams and athletes. Home crowds selectively raise and lower their noise levels based on the game situation, strategically raising the noise to disrupt the communication between members of the visiting team. Teams prefer to compete at home because they expect playing at home will gain them an edge. When home teams are on top, the superior confidence can become self-fulfilling and propel them to easy victories. However on some occasions, if home teams have not separated themselves from their perceived inferior opponents by the late stages of games, they may struggle to remain confident and the confidence of their opponents may increase. When this occurs, home teams may feel significant pressure to perform, and the competitive advantage can shift shift to the away team. The high cost of failure for home teams may also lead them to focus more on avoiding failure than striving for victory. In addition to the size of the crowd, the density and proximity to the field of play are all factors that have a large influence.
- **Referees**: In a bid to avoid creating a hostile environment, the emotional intensity of home audiences can influence decisions made by referees and officials. Some referees may be more susceptible to these influences than others. More experienced referees *should* be less biased by the impact of a large audience, which suggests that they may develop a resistance to effects of the crowd. Interestingly in 1999, the introduction of instant-replay review and the ability for teams to challenge the referees' decision on the field resulted in the home team win percentage to drop from 58.5% (1985-1998) to 56% (1998 to 2008) [16]. Despite the fact that there are seven referees officiating each game, albeit each responsible for a different area or aspect of the game, referees are continuously slandered by players, coaches, and fans for their poor officiating.
- **Travel Effect**: Travelling long distances puts athletes under physical and psychological stress. The seventeen game regular season schedule means teams will have to play a maximum of nine games away from home each year. Alternating home and away games for each individual team is not feasible, ultimately resulting in scenarios where teams will have to play multiple games on the road (away from home) for consecutive weeks. The introduction of

games being played in Europe in 2007, added a new element of planning and demands from teams for the inter-continental travel. In February 2022, the NFL announced that they were trying to further expand their reach in Europe revealing that Munich and Frankfurt will host their first NFL games in the upcoming 2022 season.

- **Familiarity**: The slight anxiety and uncertainty that everybody feels in unfamiliar territory can help shift the advantage in the way of the home team. Familiarity with the nuances of the playing surface and stadium, specific climatic conditions, wind directions, and sun position can influence HFA in multiple ways. Two studies have shown that male soccer players [17] and male ice-hockey players [18] present higher testosterone concentrations prior to home relative to away games.
- **Style of play**: Defense orientated and teams associated with low scoring games theoretically reap greater benefits from their home field advantage. For example, the Pittsburgh Steelers are historically known for their stout defense, while in recent years the Kansas City Chiefs are an explosive high scoring team. Taking the average total points in a game of 40 and 50 for these teams respectively and a HFA value of three for both teams, the Steelers would enjoy a greater benefit of playing at home ($3/40 = 7.5\%$) compared to the Chiefs ($3/50 = 6\%$).

A framework was defined to encapsulate the estimated value of home field advantage. This framework consists of finding league average HFA values and team specific HFA values accompanied by a study into the relationship between Super Bowl winning teams and their respective HFA value for that year. Referee bias was also examined by looking at the distribution of penalties called over the 22-year period.

4.2 RQ2: 4th Down, when to go for it?

A 4th Down is the last attempt the team in possession has to advance the required number of yards towards their opponent's goal line. If they fail, the other team gets the ball from that position. Teams have three options when it comes to 4th Down decision making:

- Go for it (12.69%): The offensive team can choose a regular play (run or pass). If they gain the required yards, they earn a first down and continue on as the offensive team. However, if they fail to gain the required yards, the defensive team receives possession at the spot of the tackle.
- Kick a field goal (22.21%): If the offensive team are close enough to the goal posts, they may try to kick a field goal which earns them three points. If the field goal is missed, the other team receives from possession where the ball was kicked. The field goal distance is calculated by adding seventeen yards to the current yard line the offensive team is on. Figure 4.1 provides an illustration of a 57-yard field goal from the 40-yard line.

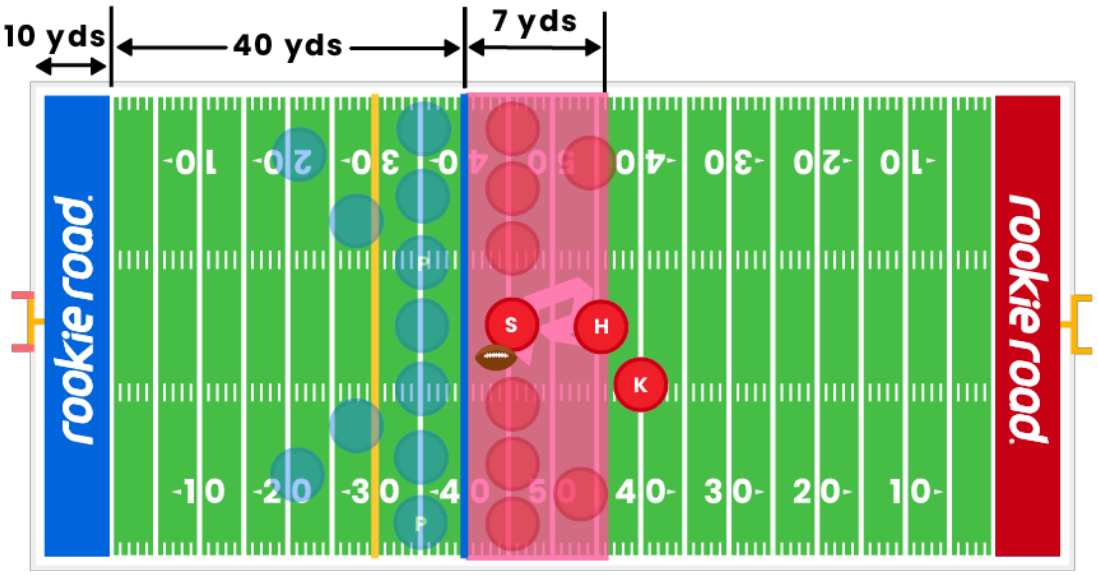


Figure 4.1: Field goal range from Rookie Road [19].

- Punt (59.64%): In most cases, the offensive team will choose to punt. A punt is the process of kicking the ball down the field as far as possible to try pin the opposing team inside their half. The average punt distance in the NFL is 45 yards [20]. Over the years teams have become more analytically driven in electing to go for it. Coaches traditionally punt or kick a field goal on 4th Downs, a convention inherited from the days when games were lower scoring and a punt was more likely to pin an opponent on his side of the field. However, modern NFL is starkly different and punts are no longer as valuable as they once were - the more freely offenses can move down the field, the less important field position becomes.

The implications of not converting a 4th Down when electing to go for it can be detrimental to the overall outcome of the game as the opposition receives possession of the ball from where the failed 4th Down occurred. 4th Down situations are becoming more common each year which has lead teams to take a closer look at the decisions they make in order to learn and make the most optimal ones for future situations. In short, making the optimal decision on 4th Down is paramount.
The analysis associated with this research question entails creating a 4th Down decision-making model while also looking at the trends associated with 4th Down situations.

4.3 RQ3: What impact has COVID-19 had on the NFL?

The COVID-19 pandemic that spread across the world at the beginning of 2020 was not only a big threat to public health, but also to the entire sports industry. Several professional sports leagues including the National Basketball Association (NBA), National Hockey League (NHL), and most professional soccer leagues in Europe took the decision to postpone or suspend their seasons in order to mitigate the spread of the disease. During an April 2020 survey conducted in the United States [21], prior to any decision taken by the NFL on the status of the 2020 season, respondents were asked on their opinion whether the NFL season should go ahead during the COVID-19 pandemic. 70% of respondents stated that the NFL should not restart to ensure players' safety. 20% of respondents expressed the view that the season should start up but allow players to choose not to play (opt in/opt out policy). 6% of respondents said to start up as planned and 5% had no opinion/didn't know (Figure 4.2).

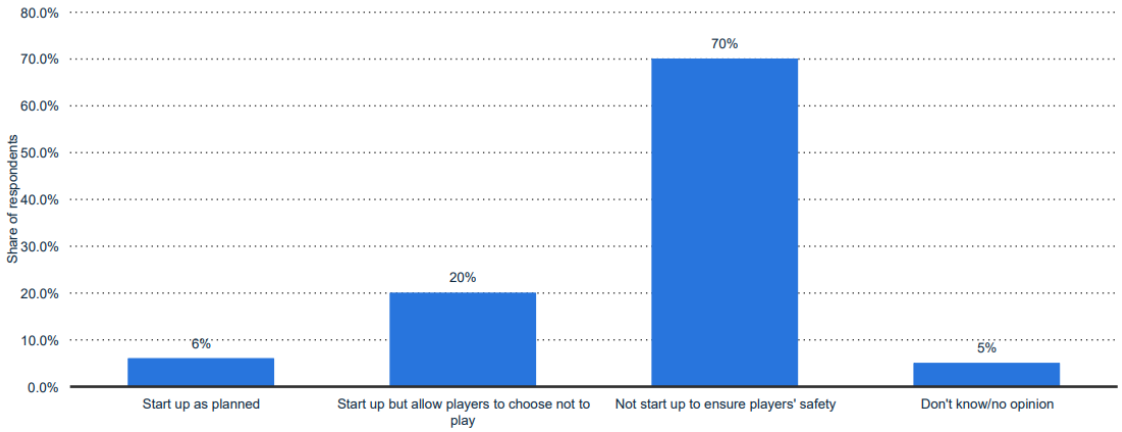


Figure 4.2: Survey conducted by Statista; April 6-8, 2020; 762 Respondents

The start of a new NFL season is signified by the annual NFL Scouting Combine (Figure 4.3). The Combine is a week-long showcase occurring every February where college football players perform physical and mental tests in front of NFL coaches, general managers, and scouts. The fundamental purpose of the Combine is to give teams a closer look at players to help make decisions during the NFL Draft held in April. The annual NFL Draft is a process where teams take turns selecting players who have just graduated from college. A team that uses an early draft pick to select a player is implicitly forecasting that this player will do well. The order of the picks is determined by the standings at the end of the season passed. The team with the worst record is granted the first pick and so on. In February 2020, the Combine went ahead as planned despite the unknown nature of the virus. However, the NFL Draft was forced to withdraw from Las Vegas amidst health concerns and headed for NFL Commissioner, Roger Goodell's basement. This signified the start of a difficult and interrupted year to come.

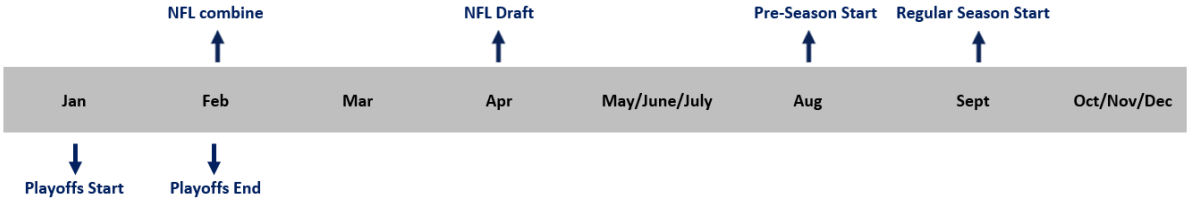


Figure 4.3: NFL Normal Season Timeline

When veterans reported to training camp on July 28, positive coronavirus cases in the United States had already reached 4.3 million [22]. The NFL was lucky to have just finished the 2019 season when COVID-19 struck, but the immediate future of the league was thrown into flux like every other business in the United States. Unlike the NBA and NHL, the NFL decided against playing the 2020 season in a bubble - where teams and players were packed into one place. The

NFL and NFL Players Association worked tirelessly together to build an environment where the players, staff, and fans would be safe while still progressing with the league. The NFL implemented a system of mitigation measures, testing, and contact tracing in an attempt to minimise COVID-19 cases. One element of this system was daily testing. NFL commissioner Roger Goodell stated that "if at any point I felt, or any of our other advisers thought, it was unsafe to continue, they were certainly prepared to take action against that" [22]. Over the course of the NFL season (August-February), 959,860 tests were administered to players, coaches, and other associated personnel. From those tests, a total of 726 people were confirmed positive COVID-19 cases - 262 players and 464 other personnel. In total, the rate of positive tests was 0.08%, substantially lower than the nationwide positive rate of 8.2% [23]. Players were also given the option to opt-out of the year which 67 players availed of including 3x Super Bowl champion Dont'a Hightower from the New England Patriots. Players who opted out received a $150,000 - $350,000 stipend for the season. Despite the rampaging global pandemic, the NFL got all 256 regular-season games played within 17 weeks in 2020.

The majority of games played in the 2020 season were without fans in attendance which giving rise to the term "ghost game". As the year went on, some teams allowed limited attendance for fans with strict mask wearing and testing protocols set in place to prevent the spread of the virus. The concluding game of the year was the Super Bowl match up between the Kansas City Chiefs and the Tampa Bay Buccaneers in Raymond James Stadium, Florida. The stadium operated at 20% capacity where 25,000 fans were in attendance. Of those 25,000 fans, 7,500 vaccinated health care workers were invited as guests of the NFL to thank them and honour them for their extraordinary service during the pandemic. The remaining 30,000 empty seats were filled by selling 30,000 cardboard cutouts to fans.

To analyse the impression that COVID-19 made on the NFL, the average points per game, favourite winning margins, HFA and 4th Down situations will be compared against previous years and the following 2021 season.

Chapter 5: Methodology

How Valuable is Home Field Advantage?

The analysis associated with this research question will aim to ascertain a league average HFA value and team specific HFA values. In tandem with these numerical representations of HFA, a case study will be carried out on the relationship of Super Bowl winning teams and their specific HFA for that year. The assumption that referees are bias towards home teams will also be investigated.

The spreadspoke_scores.csv dataset acquired from Toby Crabtree's Kaggle [12] repository will be used to calculate HFA values. This dataset contains historical information for every game played in the NFL dating back to 1999. The main information that will be used is the home/away team scores and which team is the sportsbook favourite (and by how much). The referee bias will also be investigated using the nflfastPy Play by Play dataset. The first idea is to test the community built hypothesis that home field advantage is equal to ≈ 3 points, with team specific variations of this estimate. There are many different methods used to calculate home field advantage. One method for calculating HFA for an individual team in soccer can be defined as:

$$\frac{\sum_{i=1}^{n} hGS_i - hGC_i}{n}$$

where

hGS : Goals scored at home
hGC : Goals conceded at home
n : Number of games

The direct translation to American Football for the formula above is found by replacing goals with points. However, this method is limited to only measuring how the team performs at home and will overestimate the value of HFA for top ranked teams. The team performance when playing away also needs to be considered as a comparison measure to when they're playing at home. This leads us onto the following formula which will be used to calculate HFA in this study:

$$\frac{\left(\sum_{i=1}^{n} hPS_i - hPC_i\right) - \left(\sum_{i=1}^{n} aPS_i - aPC_i\right)}{2n}$$

where

hPS : Points scored at home
hPC : Points conceded at home
aPS : Points scored away
aPC : Points conceded away
n : Number of games

Using the HFA formula we will iterate through the dataset for each team to gather team specific HFA values. The Play by Play dataset acquired from nflfastPy GitHub repository [10] will be used to analyse referee bias towards home teams. This dataset contains specific information about which team made a foul and what type of foul occurred. A penalty is a sanction called against a team for a violation of the rules, called by a referee. Referees signal penalties by tossing a bright yellow coloured flag onto the field toward or at the spot of the foul. We will look to find the ratio of home penalties to away penalties for each year to test for the presence referee bias.

4th Down, when to go for it?

The Play by Play data is the sole dataset that will be used when examining 4th Downs. To analyse the run-pass ratio when teams elect to "go for it" on 4th Down dataset will be further broken down into instances where the 'play_type' column is either 'run' or 'pass'. Through the creation of the df_run and df_pass datasets we will be able to look at how successful the run and pass decision was from each required yardage.

A 4th Down decision making tool will also be created using machine learning models. The tool tells the user to "Punt/Field Goal" or "Go for it". The input variables in the tool are:

- ydstogo: Distance required for a 1st down.
- yardline_100: The location (yard) on the field.
- game_seconds_remaining: How long is left in the game measured in seconds.
- score_differential: The difference in score from the team in possession's view.

The 4th Down data will be split into training sets and validation sets. The data will then be trained on five separate models and the model with the highest accuracy percentage will be chosen as the predictive model algorithm. The models that the data will be tested on are Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Neighbours Classifier (KNN), Decision Tree Classifier (CART), and Gaussian Naive Bayes (NB).

What impact has COVID-19 had on the NFL?

Both the PbP and historical scores datasets are used to study the impact of COVID-19 on the NFL. For the purpose of this analysis, the COVID-19 year is defined as the 2020 season. To analyse the contrast in total expected game points and total actual game points we will use the historical scores dataset. The total expected game points is defined as the over/under line set by the sportsbooks and the total actual game points is calculated from the sum of the home score and away score.
To analyse the difference in the expected winning margin for the favourite team and the actual winning margin for the favourite team, we will also use the historical scores dataset. The expected winning margin for the favourite team is taken as the spread line set by the sportsbooks. The actual winning margin for the favourite team is calculated from the favourite score minus the underdog score.
The PbP dataset will be used to study the effect COVID-19 had on home field advantage and 4th Down decision making. It would be thought that the value of home field advantage would diminish during the 2020 season due to the lack of fans at games, as the crowd noise and atmosphere are one of the main contributing factors to HFA.

Chapter 6: Results and Discussion

In this section we will look at the results of the analysis and discuss the understanding of our findings.

How Valuable is Home Field Advantage?

The first test in estimating the value of HFA was to find the league average HFA value for each year. This was calculated by averaging the home margin (points) less the away margin (points) over 2. Figure 6.1 plots the league average HFA value versus year (1999-2021). We observe a pattern where HFA has decreased over time. The decline in HFA is very important to the sports gambling industry where point spreads are assigned to each game. In recent years, it appears that the regression rate of HFA is increasing, reaching an all-time low point of -0.002 in 2019. The steep regression in the value of home field advantage across the league indicates that in the coming years, the slight edge of playing at home might be totally nullified. One of the reasons behind this change is that there have been massive strides taken in sports science, technology, and logistics that have greatly reduced the negative impact of long travel times. Namely, the sleeping pattern and diet of players are being closely watched by teams of experts to ensure the most optimal performance on game day.

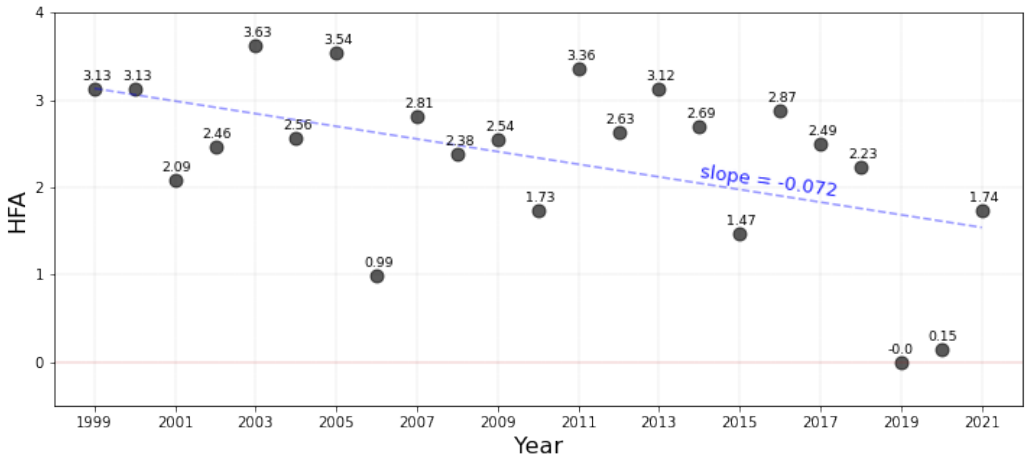


Figure 6.1: League average HFA per year

It is clear that a team's primary objective is to win games rather than maximising the score differential. We also considered the home team win rate as a supporting measure of home field advantage. Figure 6.2 reinforces our two findings that home field advantage is present but appears to be diminishing. The drop off in home team win percentage from 2018 to 2019 (59.7% to 52.1%) coincides with the drop off in our HFA values from 2018 to 2019 (2.23 to -0.002).

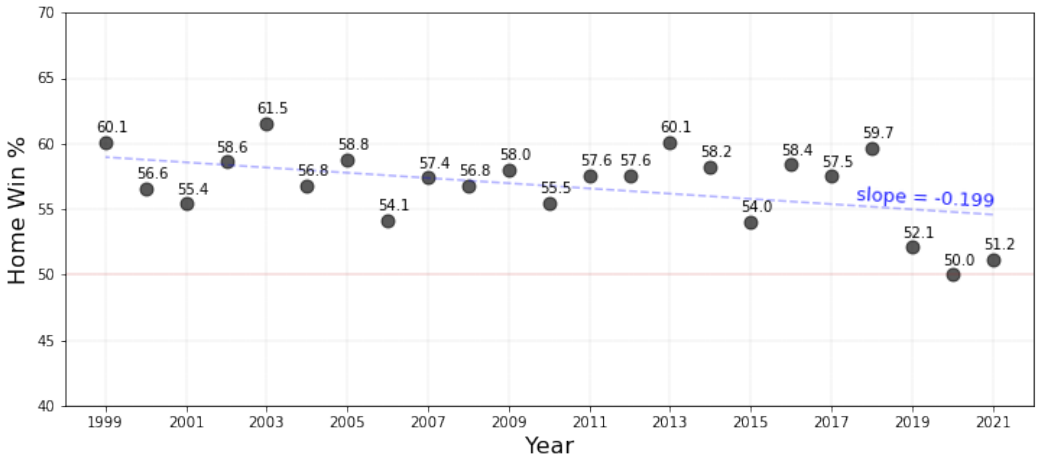


Figure 6.2: Home team win rate

Figure 6.3 shows the average HFA for each team in the NFL dating back to 1999. The teams possessing the highest HFA are the Minnesota Vikings, Green Bay Packers, and Seattle Seahawks. The Minnesota Vikings are surprising as league leaders given their mediocre 51.04% total win percentage. Hailing from the same division, the Green Bay Packers are considered to be one of the top franchises in the NFL. The combination of the cold weather in Northern USA and a loyal fan base was no surprise to see the Packers high up on the list. The Seattle Seahawks can attribute their high HFA value to their crowd who carry "The 12th Man" name. Perhaps most surprising are the New England Patriots, presenting a less-than-league-average HFA value of 2.12. Throughout the analysed period, the New England Patriots were a dominant force in the NFL, hoisting the Lombardi trophy on six occasions with Quarterback Tom Brady (2000-2019) and Head Coach Bill Belichick at the helm (2000-present). The Patriots won an astonishing 77.67% of home games through this period by an average margin of 9.67 points. This dominance also carries over to their away games - winning 63.77% of the time by an average margin of 5.42 points. Though, the overall strength of the team must be taken out of consideration when analysing HFA, thus the HFA value perceives the Patriots to be a weak team at home when in fact they were one of the most dominant home teams from 1999-2021.

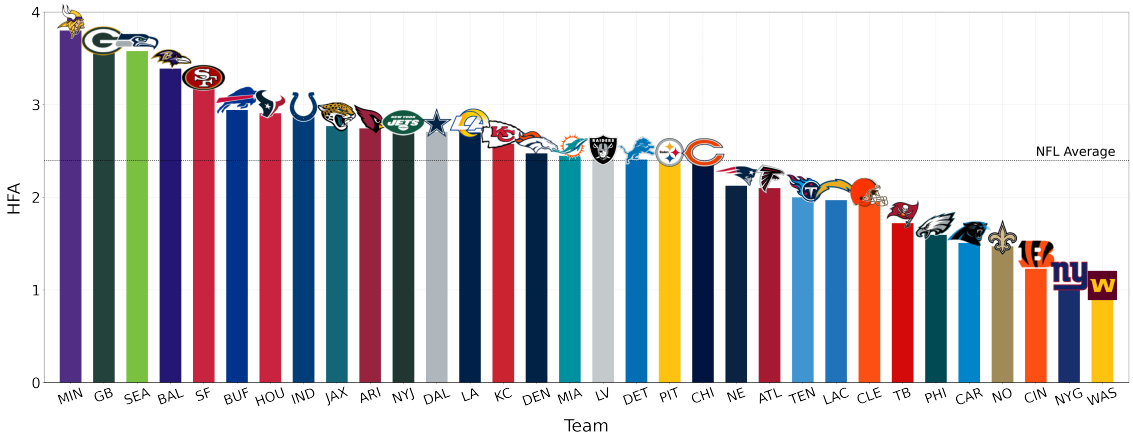


Figure 6.3: Team specific HFA (since 1999)

To further investigate the effect of home field advantage, Figure 6.4 represents the bias shown towards home teams in relation to referee decisions. 51,677 penalties called by referees from 2001-2021 were analysed. On average, away teams were penalised 2.49% more often than home teams. This figure remained relatively consistent over the analysed period (sd = 1.03%).

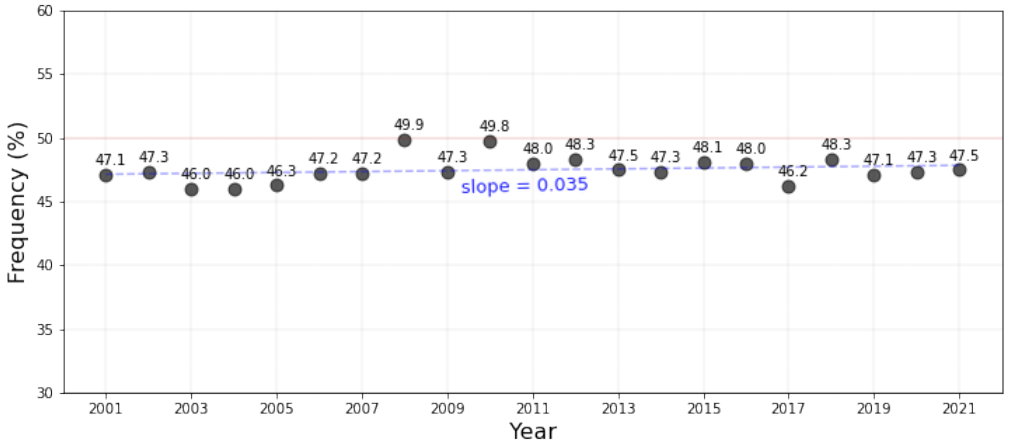


Figure 6.4: Home team penalties conceded

Some penalties are more common and subjective than others. Table 6.1 shows the top 5 penalties called in the NFL. Offensive false start penalties are the most common penalties in the NFL and account for 26.79% of all penalties called. An offensive false start is when an offensive player moves in a way that simulates that the play is starting prior to the ball being snapped [24]. Perhaps the most controversial and subjective penalty called by referees is defensive pass interference (DPI). DPI is when any act by a player past the line of scrimmage significantly hinders an offensive player's opportunity to catch the ball [25]. Referees are 4.33% more inclined to penalise the away team for DPI which supports the overall value for home team referee bias of 2.49%. The uncertainty around how much contact the defender is allowed to make before it is deemed interference ultimately leads to referees calling penalties on their own personal interpretation of the rule, leaving them open to be potentially swayed by overwhelming crowd pressure and noise.

Penalty Type	Frequency	Home team freq.	Away team freq.
Offensive False Start	26.79%	48.63%	51.37%
Offensive Holding	15.77%	51.47%	48.53%
Defensive Pass Interference	9.41%	45.67%	54.33%
Defensive Offside	7.48%	50.66%	49.34%
Defensive Holding	5.60%	47.39%	52.61%

Table 6.1: Top 5 penalties

Our final test for estimating the value of home field advantage is to study its relevance to Super Bowl winning teams. Figure 6.5 shows that 5 of the last 23 Super bowl winners had a negative HFA value. Additionally, only 39.13% of Super Bowl winning teams had a HFA value greater than the league average for that year.

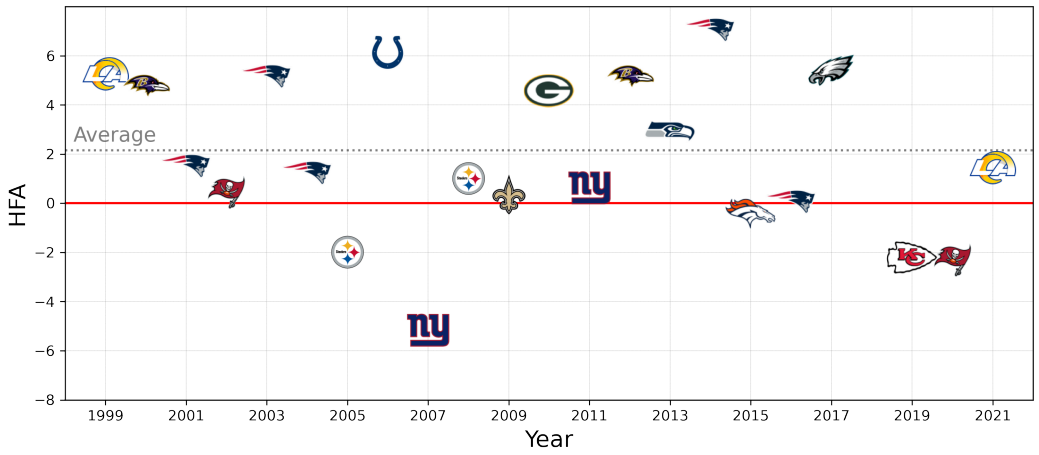


Figure 6.5: HFA for Super Bowl winning teams

4th Down, when to go for it?

Figure 6.6 shows that the number of 4th Down occurrences are decreasing year by year. However, through the power of new analytical tools and critical thinking coaches, the rate at which teams are "going for it" (run or pass) has increased substantially since 2017 (Figure 6.7). The most recent season (2021) saw teams go for it 15.76% of the time. This is a 120.24% increase from 10 years ago.

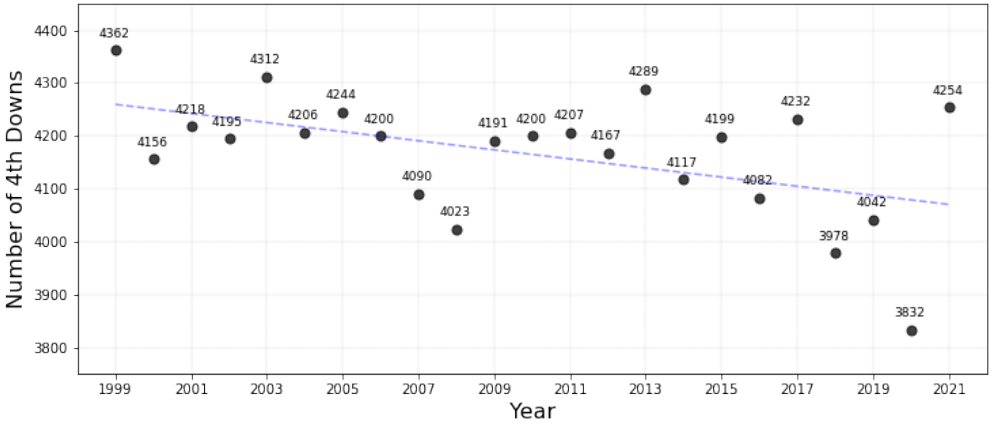


Figure 6.6: 4th Downs Per Year

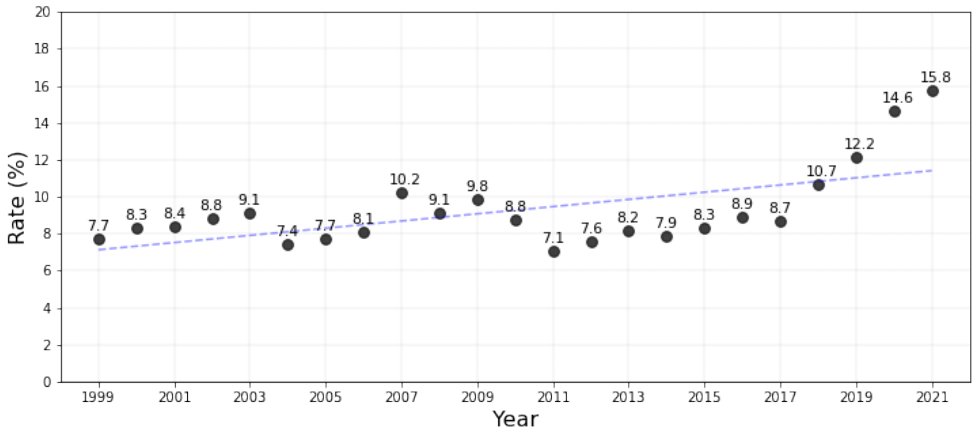


Figure 6.7: Go for it rate on 4th Down

Teams have two choices when going for it on 4th Down - run or pass. Teams elect to pass 54% of the time and run 46% of the time. By electing to pass, teams are putting the result of the play in the hands of their Quarterback, usually their best player. However, on short yardage it makes sense to use more of a "brute force" technique and run for the required short distance. Figures 6.8 and 6.9 show the "go for it" rate and success rate when teams elect to go for it on 4th-1, 4th-2 and 4th-3. Teams are electing to go for it on 4th-1 at an unquestionably high rate (64.12% in 2021) and this rate looks like it will continue trending upwards for the coming 2022 season. Conversely, the success rate on 4th-1 has not shown the same increase as the "go for it" rate. Perhaps teams now have a larger sample size to work with when analysing the opposition teams tendencies on 4th-short (yards 1-3).

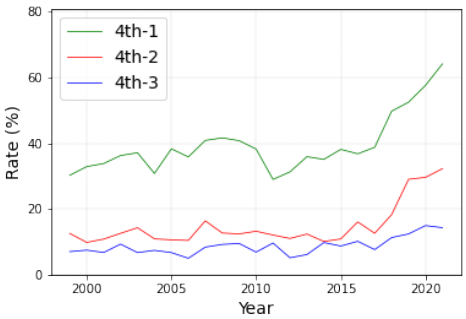


Figure 6.8: "Go for it" rate

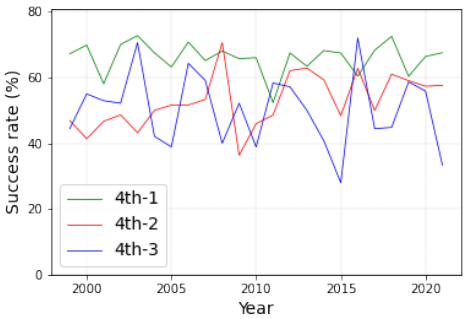


Figure 6.9: "Go for it" success rate

Figure 6.10 shows the conversion rates for each required distance when a team elects to pass on 4th Down. There is a steady decrease in both attempts and the conversion rate as the yardage increases from 1 to 5. The average conversion rate for passing on 4th Down is 43.72% for all yardages. This number jumps to a near 50:50 split (49.1%) when selectively looking at yards 1 to 5.

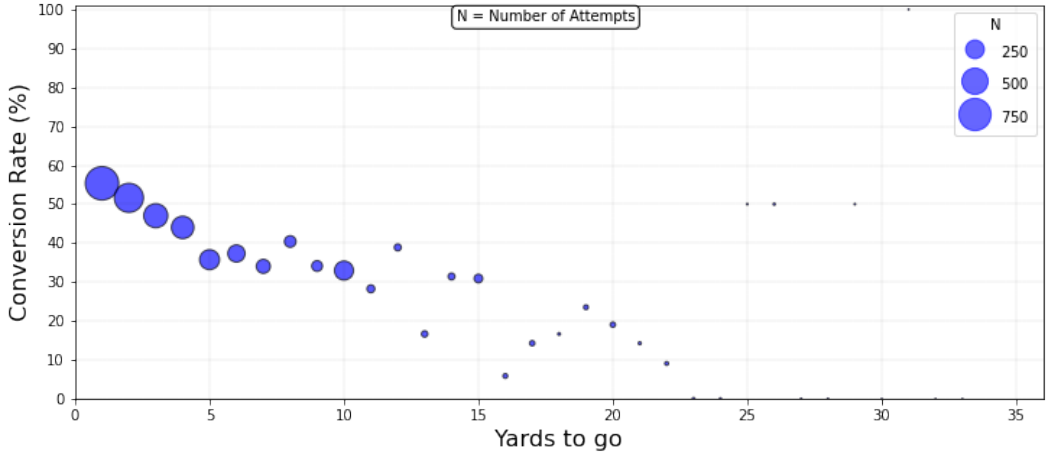


Figure 6.10: 4th Down Pass Conversion Rate

Figure 6.11 shows the conversion rates for each required distance when a team elects to run on 4th Down. It is no surprise that running on 4th-1 takes up 87.21% of all run attempts on 4th Down. Running on 4th-1 has a higher success rate (69.54%) than passing (55.35%). It is thought that the bulk of these 4th-1 run plays are attributed to the unique Quarterback Sneak play. A Quarterback sneak (QB sneak) is a play in American Football in which the quarterback dives ahead with the football while the offensive line surges forward [26]. QB sneaks can be very effective in short yardage situations.

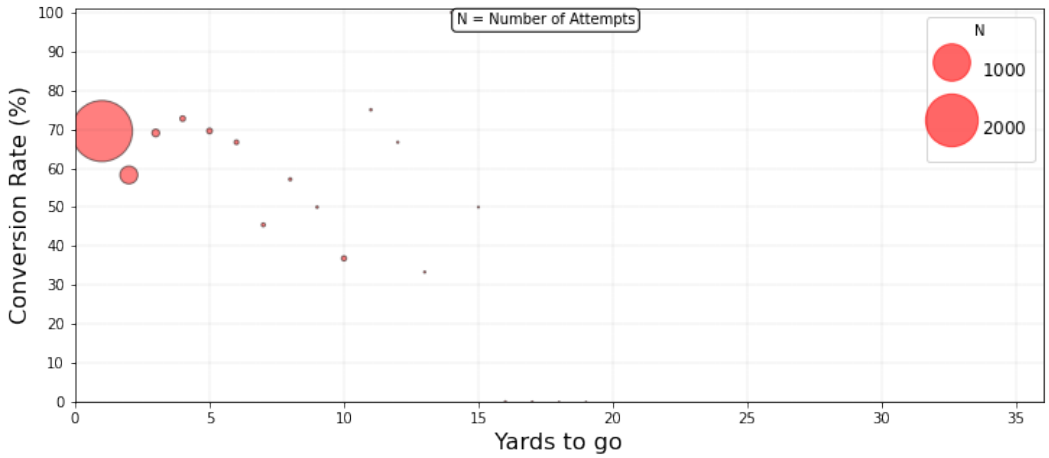


Figure 6.11: 4th Down Run Conversion Rate

From Figures 6.10 and 6.11 we can conduct a very simple decision tree (Figure 6.12) for deciding between running or passing on 4th Down when electing to go for it.

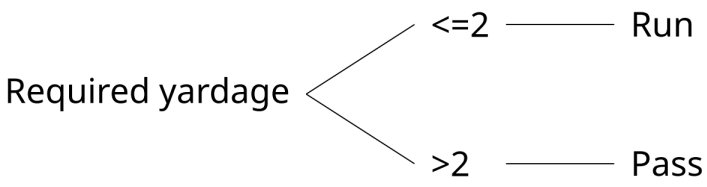


Figure 6.12: Go for it decision tree

Figure 6.13 displays the accuracy scores ± standard deviations obtained when testing the 4th Down data on the following five models: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Neighbours Classifier (KNN), Decision Tree Classifier (CART), and Gaussian Naive Bayes (NB).

```
LR: 91.12 ±0.1
NB: 90.9 ±0.17
LDA: 91.06 ±0.05
KNN: 91.45 ±0.25
CART: 90.48 ±0.4
```

Figure 6.13: Model Performances (%)

The K-Nearest Neighbours algorithm (KNN) was selected as our model for predicting 4th Down decision making as it exhibits the highest accuracy score from our model testing (91.45%). The next step was to choose the optimum value for k. Choosing an optimum value for k can be difficult as a small value means that dataset noise will have a higher influence on the result. Choosing a value for k that is too large will under-fit the model causing the error rate to rise. Figures 6.14 and 6.15 show that when k=9 we minimise the error rate to 0.084 while maximising the accuracy to 91.45%.

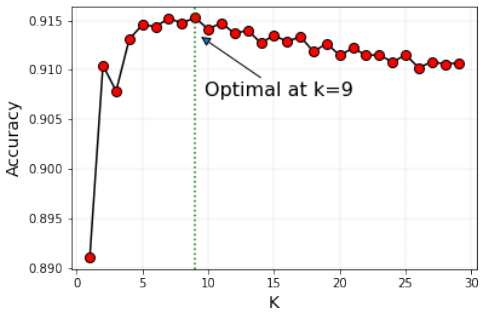


Figure 6.14: Accuracy Testing

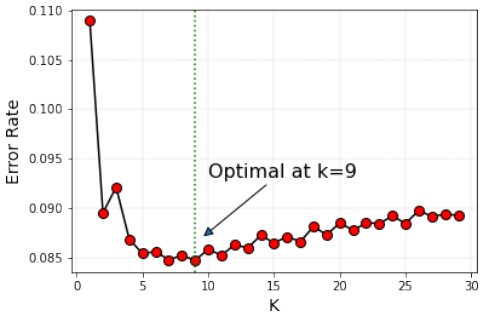


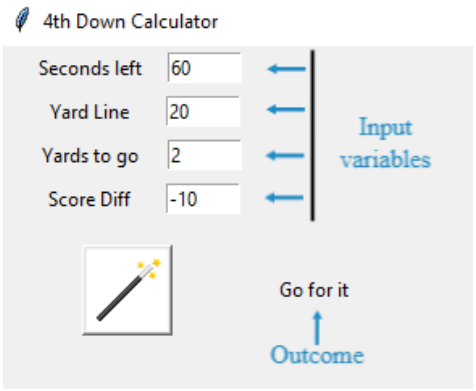
Figure 6.15: Error Testing

In order to make it easy to test the model on sample scenarios, a Graphical User Interface (GUI) was created as a platform for the model to run on. The tkinter Python library [27] was used to create the GUI to run our calculations on (Figure 8.6). Some sample scenarios (Table 6.2) were created to test how the model performs for different input values.

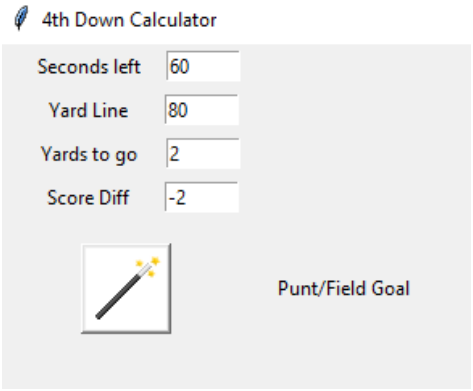
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Seconds Left	60	60	120	600
Yard Line	20	80	87	20
Yards to Go	2	2	1	2
Score Differential	-10	-2	6	25
Model Decision	Go for it	Punt/Field Goal	Punt/Field Goal	Punt/Field Goal

Table 6.2: 4th Down Scenarios

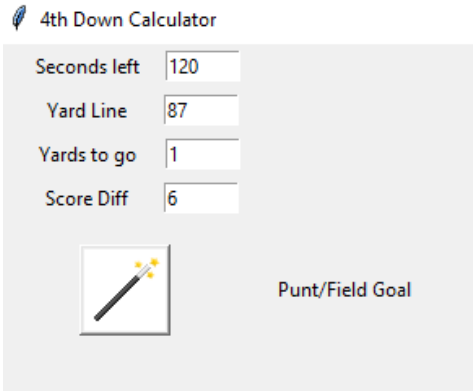
In scenario 1 (Figure 6.16a), the team in possession is already in a difficult situation and the probability of winning is relatively low. It is too far for a field goal and punting would guarantee a win for the opposition team. Go for it is the correct decision.
Conversely, in scenario 2 (Figure 6.16b) the team in possession is in a good position to win the game. A field goal from the 20-yard line is 37 yards away from the posts. This is considered a relatively easy kick. Following a successful field goal (3 points) the team in possession will be leading by 1 point. There is no reason to punt here. Punt/Field Goal is the correct decision.
Similar to the previous scenario, in scenario 3 (Figure 6.16c) the team in possession is in a good position to win the game. A field goal from the 13-yard line is 30 yards away from the post. This is considered a very easy kick. Following a successful field, goal the team in possession will be leading by 9 points. This means the opposing team will have to score twice in order overturn this 9 point differential. There is no reason to punt here. Punt/Field Goal is the correct decision.
In scenario 4 (Figure 6.16d), the team in possession is in a very good position to win the game. It is too far for a field goal attempt. The team in possession is winning by 25 points which implies they are significantly better than the opposition. However, the risk/reward in going for it in this situation is not worth it. Punt/Field Goal is the correct decision.



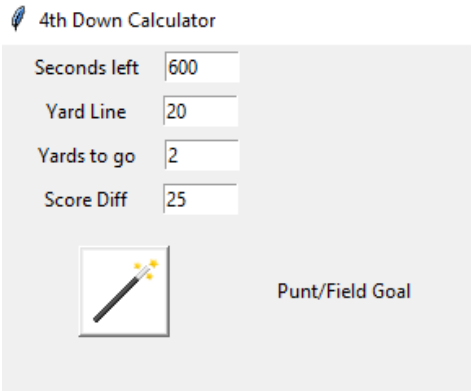
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3



(d) Scenario 4

Figure 6.16: Model Examples

A video demonstration of the model working for scenario 1 can be found in section 8.1.

What impact has COVID-19 had on the NFL?

Figure 6.17 shows the distribution of the predicted margin for favourite teams from 2019-2021. The predicted margin of victory was defined as the spread set by sportbooks. Figure 6.18 illustrates the distribution of the actual margin for the favoured team. Interestingly, there were more outliers for the negative y-values of the boxplot indicating that during the COVID-19 season, there was a notable increase in the number of large margin victories by teams carrying the underdog tag. The biggest upset of the 2020 season occurred in week 9 when the New Orleans Saints beat the reigning Super Bowl champions Tampa Bay Buccaneers 38-3. The Buccaneers entered the game 4 point favourites (win probability = 65.77%).

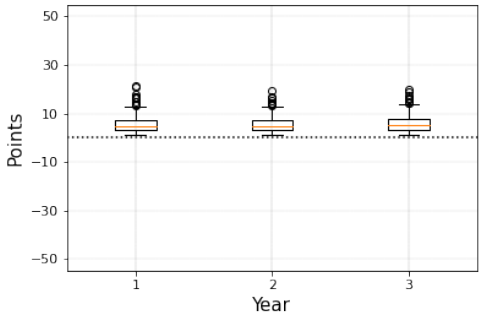


Figure 6.17: Predicted margin

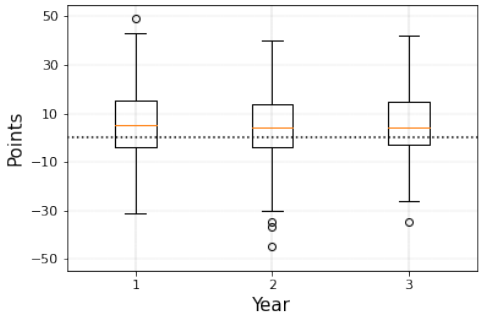


Figure 6.18: Actual margin

Figure 6.19 shows the distribution of the predicted total points from 2019-2021. The predicted total points was defined as the over/under line set by sportbooks. Figure 6.20 shows the distribution of the actual total points. There was a significant rise in the average total number of points per game from 2019 (45.70) to 2020 (49.5). This spike returned back to the norm in 2021 (46.06).

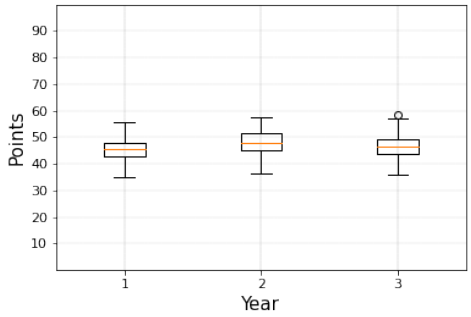


Figure 6.19: Predicted points

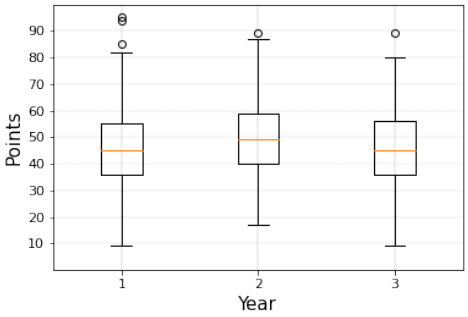


Figure 6.20: Actual points

Evidently, the COVID-19 season showed an increase in the amount of upsets caused by underdog teams and an overall increase in the total amount of points scored in each game. It is important to look at whether the favourite or underdog team were the biggest drivers in the increase of total points. Table 6.3 shows that the favourite team had a slight 6.9% increase in average scoring from 2019-2020 but the underdog team exhibited an 11.8% increase in average scoring from 2019-2020. The greater increase in average scoring for the underdog shows that they benefited more during the COVID-19 season in terms of point scoring averages, consequently causing more upsets.

	2019	**2020**	**2021**
Total points (avg)	45.7	49.5	46.07
Favourite points (avg)	25.2	26.94	25.68
Underdog points (avg)	20.18	22.56	20.39

Table 6.3: Scoring Distribution

The value of home field advantage was impacted significantly in the 2020 season primarily due to the limited fans in attendance. Home field advantage was present, albeit quite minimal. Interestingly, Figure 6.21 shows that HFA in 2019 (the year prior to COVID-19) reached negative value of -0.002 implying there was an average home field *disadvantage* during that year. HFA was valued at 0.15 in 2020 and jumped up to 1.74 the following season. Additionally, by considering the home/away win ratio as a supporting measure for HFA we can see that there was an even 50:50 split in the 2020 season (Figure 6.2). This supports the idea that HFA was minimal during the 2020 season.

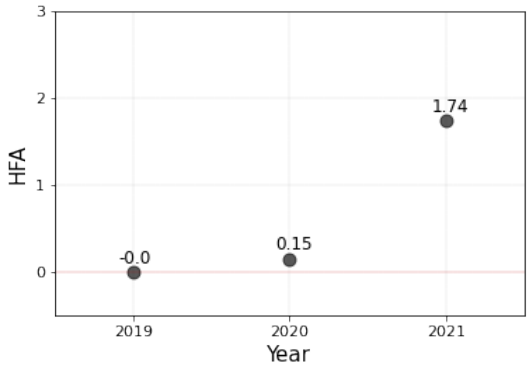


Figure 6.21: Home Field Advantage

The observed increase in total points during the 2020 season implies that the offensive team was able to score with more freedom in comparison to previous years. Subsequently, it is no surprise to see that there was a 5.19% (4042 to 3832) decrease in the number of 4th Down situations from 2019 to 2020. Figure 6.22 shows how the number of 4th Down situations fluctuated significantly from 2019-2021. Despite the decrease in the number of 4th Downs, the factors influencing the 2020 season had no effect on the go for it rate on 4th Down (Figure 6.23) which continued on its upward trend.

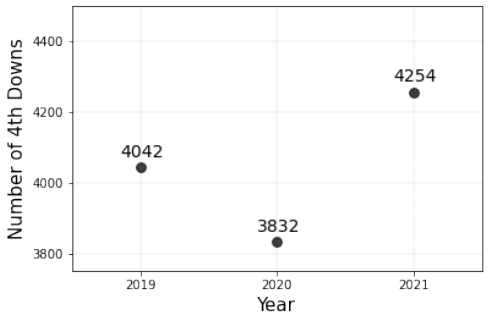


Figure 6.22: 4th Down Situations

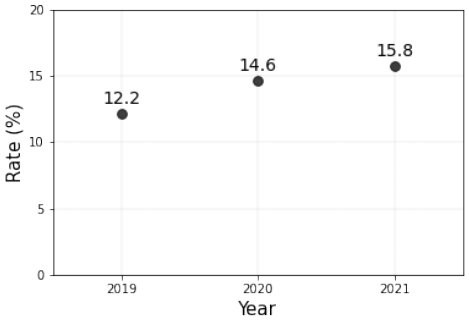


Figure 6.23: 4th Down Go For It Rate

Chapter 7: **Conclusions**

In this section we will conclude our analysis on our three main research questions.

7.1 Main Findings

The analysis presented on home field advantage showed that teams do gain an edge when playing at home. The league average home field advantage value as of the 2021 season is 1.74 and this value is declining year by year. However, this value is not fixed and deviates with respect to different teams. The Minnesota Vikings, Green Bay Packers, and Seattle Seahawks exhibited the highest values for HFA while the Cincinnati Bengals, New York Giants, and Washington Commanders presented the lowest. The relationship between Super Bowl winning teams and a strong HFA value for that year is minimal. The investigation into the officiating in the NFL showed that referees tend to favour the home team by penalising the away team 2.49% more often, on average. Considering the above, it can be concluded that home field advantage does exist in the NFL, albeit team-deviating and diminishing each season.

The investigation on the 4th Downs observed that 4th Down situations are increasing each year. The rate at which teams are "going for it" is also increasing - 15.8% as of the 2021 season. On the analysed 4th-1 situations, electing to run presented a 69.54% success rate, greater than the 55.35% success rate when teams elected to pass. The K-Nearest Neighbours model created chooses the most optimal decision to make on 4th Down when considering defined parameters, to an accuracy score of 91.45%.

COVID-19 had a significant impact on the play style and outcomes of games throughout the 2020 season. The average number of points per game increased by 8.31% from the year previous. The underdog team had a greater percentage increase in points scored (11.8%) than the favourite team (6.9%). Consequently, there was a notable increase in the number of large margin victories by the underdog team. The value of home field advantage was notably low during the 2020 season (0.15), perhaps due to the limited fan attendance at games. The number of 4th Down situations decreased in the 2020 season by 5.19% from the year previous. However, the rate at which teams elected to "go for it" (15.8%) stayed consistent, following the upward trend from recent seasons.

7.2 Limitations

One of the main limitations of this project is the sample size of data. The data used for this project dates back to 1999. Looking at seasons pre 1999 would be likely to have an impact on the overall findings and conclusions illustrated here. Furthermore, the sample size is greatly hindered given the structure of the NFL season. The new NFL schedule consists of teams playing a maximum of 22 games per season. In comparison to other American sports, the NFL plays a small number of games each season. For example, Major League Baseball (MLB) schedules 162 games for each of the 30 teams per year. The National Basketball Association (NBA) schedules 82 games for each of the 30 teams per year. Having a larger sample size of data would create more accurate findings. An assumption was made to the related 4th Down data where all occurrences of 4th Downs where the team in possession had a win probability of less than 5% or greater than 95% were removed from the dataset. This was done in an attempt to remove situations where teams acted more aggressively given the game state. This threshold was personally defined and is subjective to the person analysing the data. There are also some limitations to the 4th Down model created. Considering more variables such as "timeouts remaining" could provide a more accurate prediction. Additionally, the model currently includes "field goal" and "punt" as the same output. Splitting these into two separate outputs would provide another dimension to the predictive tool.

7.3 Future Work

The scope of this project was confined to professional American Football in the NFL. Other leagues that play American Football exist such as the XFL (Alternative American Football League) and the NCAAF (National Collegiate Athletic Association for Football). SportsDataIO [28] provides an API for acquiring detailed data on games played in the NCAAF. A suggestion for future research would be to apply the methods and tools implemented in this project to the XFL and the NCAAF and to compare the results observed across each league. There is also ample capacity for further exploration into the trends that surface when analysing 4th Downs. The current analysis of a team's decision on 4th Down which is split into run or pass could be further divided into what formation the team used and the direction they ran or passed. The strength of the kicker and punter could be considered in the 4th Down model which would provide different outputs for different teams based on the quality of the kicker and punter.

Chapter 8: Appendices

8.1 Additional Work

A video demonstration of the 4th Down predictive model:

- https://youtu.be/Iepcz_DfH3cL

Home teams win on average 56.78% of games. The home team win percentage remains consistent through the years exhibiting a standard deviation of 2.89 (Figure 8.1).

Season	Number of Games	Home Wins	Away Wins	Ties	Home Win %
2021	281	144	136	1	51.25%
2020	268	134	133	1	50.00%
2019	261	136	124	1	52.11%
2018	263	157	104	2	59.70%
2017	261	150	111	0	57.47%
2016	262	153	108	1	58.40%
2015	263	142	121	0	53.99%
2014	263	153	109	1	58.17%
2013	263	158	104	1	60.08%
2012	264	152	111	1	57.58%
2011	264	152	112	0	57.58%
2010	263	146	117	0	55.51%
2009	264	153	111	0	57.95%
2008	264	150	113	1	56.82%
2007	265	152	113	0	57.36%
2006	266	144	122	0	54.14%
2005	262	154	108	0	58.78%
2004	266	151	115	0	56.77%
2003	265	163	102	0	61.51%
2002	266	156	109	1	58.65%
2001	258	143	115	0	55.43%
2000	258	146	112	0	56.59%
1999	258	155	103	0	60.08%

Table 8.1: Home team win rates

The average home team win percentage from 1999-2021.

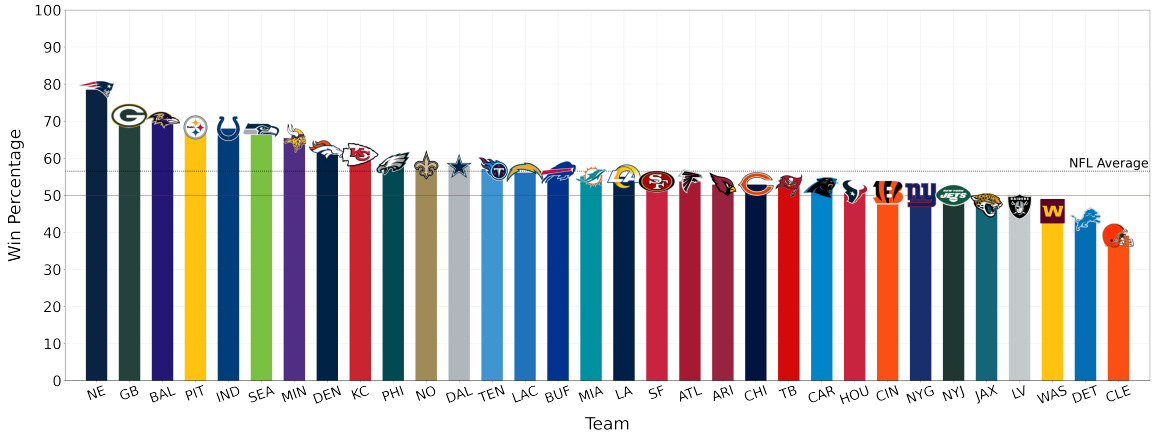


Figure 8.1: Home Team Win Percentage

Table 8.2 shows the structure of dataset used to calculate team specific HFA.
The formula used can be found in section 5.

Team	Home Record	Home Margin	Away Record	Away Margin	HFA
MIN	124-66-0	4.28	72-120-2	-3.32	3.80
GB	143-55-2	7.45	102-96-0	0.28	3.59
SEA	133-64-0	5.83	91-108-1	-1.32	3.58
BAL	136-57-0	7.67	97-106-0	0.89	3.39
SF	102-88-1	1.84	76-119-0	-4.55	3.19
BUF	103-85-0	2.13	72-117-0	-3.75	2.94
HOU	85-82-0	0.28	58-106-0	-5.54	2.91
IND	135-64-0	5.62	108-90-0	-0.13	2.88
JAX	89-99-0	0.33	60-129-0	-5.21	2.77
ARI	100-87-2	-0.39	67-123-0	-5.87	2.74
NYJ	93-93-0	0.30	75-120-0	-5.12	2.71
DAL	110-80-0	3.52	87-104-0	-1.89	2.71
LA	108-89-0	2.15	84-107-1	-3.24	2.70
KC	122-77-0	4.33	88-101-0	-0.82	2.58
DEN	121-75-0	3.41	86-102-0	-1.53	2.47
MIA	102-86-0	0.55	77-111-0	-4.34	2.45
LV	89-102-0	-1.17	66-121-0	-5.99	2.41
DET	80-104-0	-1.90	50-137-2	-6.71	2.41
PIT	137-61-2	6.26	110-85-1	1.49	2.38
CHI	101-90-0	1.37	79-108-0	-3.39	2.38
NE	167-48-0	9.67	125-71-0	5.42	2.12
ATL	102-88-0	1.30	84-108-1	-2.90	2.10
TEN	109-82-0	1.58	95-101-0	-2.42	2.00
LAC	106-84-0	3.81	85-106-0	-0.13	1.97
CLE	72-112-1	-3.58	49-138-0	-7.49	1.95
TB	99-92-0	1.66	83-110-0	-1.78	1.72
PHI	116-82-1	4.40	106-90-1	1.22	1.59
CAR	99-91-0	0.93	84-109-1	-2.09	1.51
NO	113-81-0	3.52	104-90-0	0.58	1.47
CIN	96-91-3	-0.69	70-119-1	-3.27	1.29
NYG	95-95-0	-0.04	90-105-0	-2.14	1.05
WAS	86-102-0	-1.70	72-116-1	-3.59	0.95

Table 8.2: Team Specific HFA

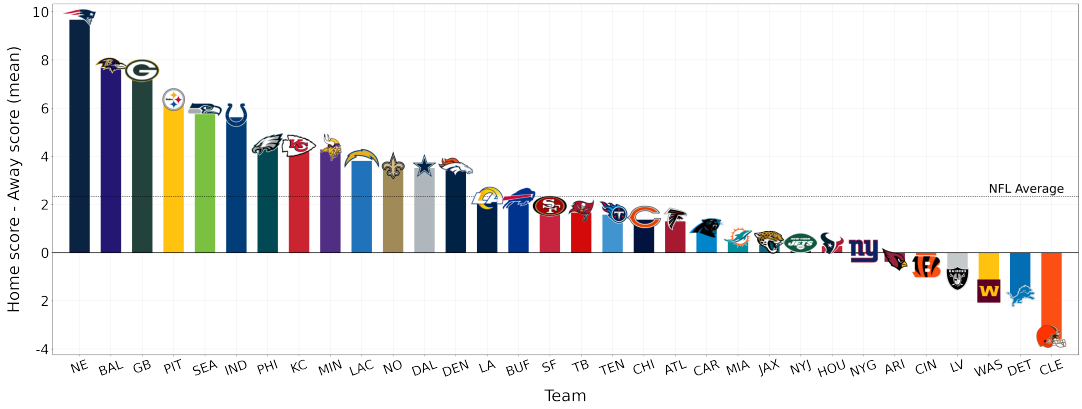


Figure 8.2: Home Team Average Point differential

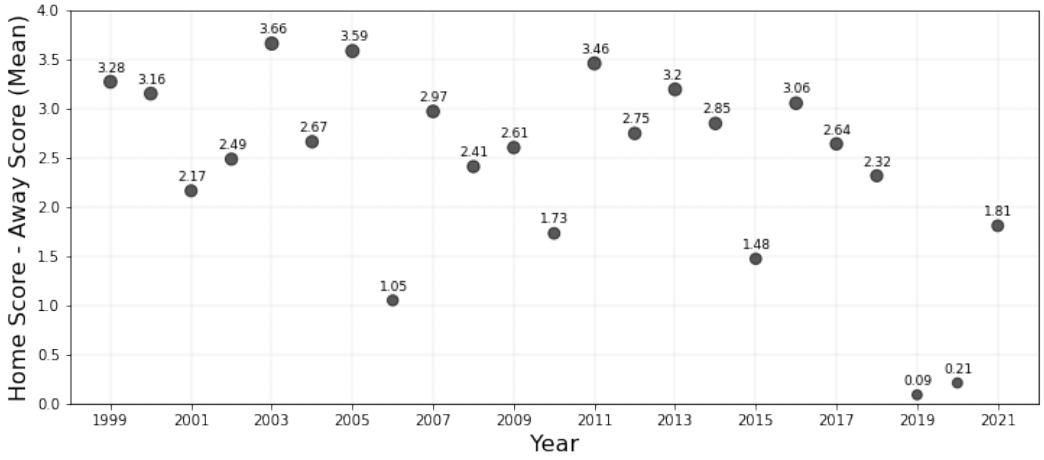


Figure 8.3: Home Team League Average Point differential

Figure 8.4 shows the mapping of the spread set by sportsbooks to a numerical win percentage value.

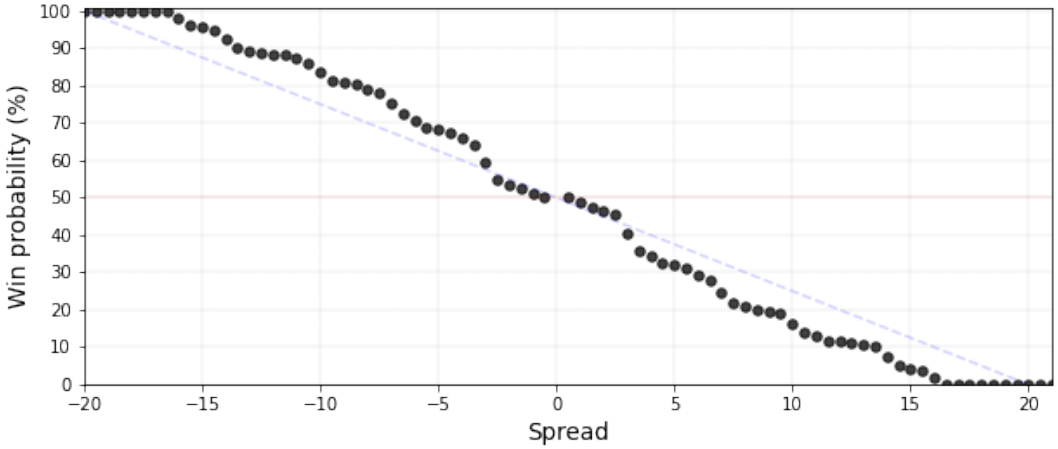


Figure 8.4: Spread to win percentage mapping

Figure 8.5 shows the lines set by a major sportsbook (bet365) for the upcoming pre-season game in August 2022 between the Las Vegas Raiders and the Jacksonville Jaguars. The spread indicates which team is the favourite and by how much. In this example, the Las Vegas Raiders are the favourite and the sportsbook expects them to win by 1 point. The sportsbook also expects there to be ≈ 33.5 total points in this game.

Game Lines

	LV Raiders	JAX Jaguars
Spread	-1.0 1.90	+1.0 1.90
Total	O 33.5 1.90	U 33.5 1.90

Figure 8.5: Example of Spread and Total Lines set by bet365 [29]

GUI created for running 4th Down Predictive Tool.

Figure 8.6: Model GUI

8.2 Additional Information

Game

NFL games are 1 hour long, divided into 4 quarters of 15 minutes each. The clock is stopped frequently - after an incomplete pass or any play that ends up out of bounds. Each team is allowed 3 timeouts per half that they can use at their own discretion. NFL games can exceed three hours in duration.

The entire NFL field is a large rectangle 360 feet long and 160 feet wide. The field is demarcated by hash marks at each yard and horizontal lines across the field every 10 yards. At every 10 yards, the yard number is labeled on the line. The labels start at 10 and increase as they get further away from the end zones, meeting in middle at 50. Figure 8.7 illustrates the markings and dimensions of the field of play.

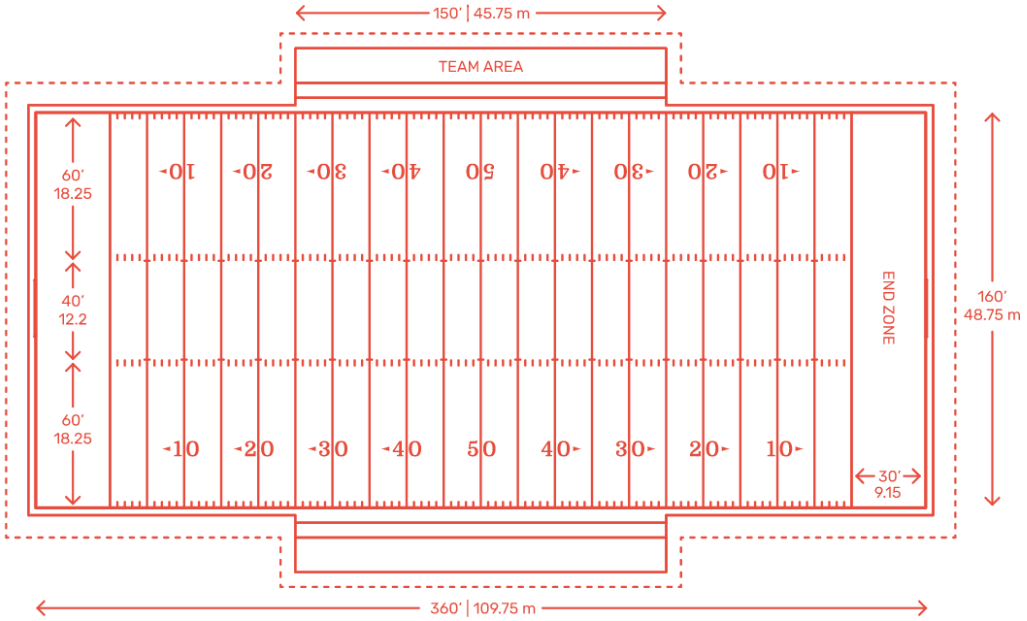


Figure 8.7: NFL Field Illustration [30]

There are four ways to score points in the NFL:

- Touchdown: A touchdown is earned when the ball either crosses the other team's goal-line while in the possession of a player, or a player catches and secures the ball while in the opposing team's end zone. A touchdown is worth 6 points.
- Try play: After a touchdown, the offensive team has the choice of kicking a field goal from the 15-yard line to earn 1 point or attempt to reach the end-zone in the same ways as a touchdown with the play starting from the 2-yard line to earn 2 points.
- Field Goal: A field goal is When a player (usually a designated kicker) kicks the ball through the uprights earning 3 points.
- Safety: A safety occurs when the offensive team is tackled or touched down by a defender in their own end zone. The play scores 2 points for the defensive team.

Teams

Each team can carry a maximum of 53 total players on their roster however, only 48 of these can be active on game day. There can only be 11 players on the field from each team at any one time. Each team is further broken into three task-specific sub-teams:

- Offense: The team in possession of the ball, trying to score. This team consists of positions such as the Quarterback (QB), Running Back (RB) and Wide Receiver (WR).
- Defense: The team trying to prevent the other offense scoring. This team consists of positions such as Linebacker (LB), Cornerback (CB), and Safety (S).
- Special Teams: This team plays all kicking situations (field goals and punts). Some players only play on special teams such as the Punter (P) and Kicker (K). The other players play on Special Teams and Offense or Defense.

The NFL consists of 32 teams broken up into 2 conferences - American Football Conference (AFC) and National Football Conference (NFC). These 2 conferences are then broken up into 4 divisions (North, South, East, West) each consisting of 4 teams. Figure 8.8 shows the breakdown of teams into their respective conferences and divisions.

Figure 8.8: Conference and division breakdown [31]

Schedule

In 2021, the NFL moved from a 16 game season to a 17 game season. Table 8.3 illustrates the breakdown of schedule history over recent years.

Year	Weeks	No. Teams	Reg. Season Games	Post Season Games	Total Games
2021	17	32	272	13	285
2020	16	32	256	13	269
2019	16	32	256	11	267
2018	16	32	256	11	267
2017	16	32	256	11	267
2016	16	32	256	11	267
2015	16	32	256	11	267
2014	16	32	256	11	267
2013	16	32	256	11	267
2012	16	32	256	11	267
2011	16	32	256	11	267
2010	16	32	256	11	267
2009	16	32	256	11	267
2008	16	32	256	11	267
2007	16	32	256	11	267
2006	16	32	256	11	267
2005	16	32	256	11	267
2004	16	32	256	11	267
2003	16	32	256	11	267
2002	16	32	256	11	267
2001	16	31	248	11	259
2000	16	31	248	11	259
1999	16	31	248	11	259

Table 8.3: Schedule History Breakdown

In this new updated schedule, each team plays 17 regular season games with one bye week. The bye week is a specified week when a team doesn't play a game. Teams alternate seasons where they host nine regular season games or eight regular season games. The breakdown of schedule is listed below:

- Six games against divisional opponents — two games per team, one at home and one away.
- Four games against teams from a division within its conference — two games at home and two away.
- Four games against teams from a division in the other conference — two games at home and two away.
- Two games against teams from the two remaining divisions in its own conference — one game at home and away. Matchups are based on division ranking from the previous season.
- The new 17th game is an additional game against a non-conference opponent from a division that the team is not scheduled to play. Matchups are based on division ranking from the previous season.

Once the regular season is finished, 7 teams from each conference qualify for the post season: the 4 division winners plus the three best runners up in each division. Each conference conducts a seeded single-elimination knockout tournament to decide the conference winners. The team with the best regular season record gets a bye into the second round of the elimination tournament. The separate stages of the elimination tournament are: wildcard round, divisional round, and conference round. The winners of the conference round play each other in the Super Bowl to decide the league champion.

Overview of Key Parameters

AFC: American Football Conference. One of two conferences in the NFL consisting of 16 teams.

Bye Week: An off week where no game occurs. Each team is scheduled one per season.

Conference: Third Round of playoffs.

Divisional: Second round of playoffs.

End zone: The scoring area on the field.

Ghost Games: Games without fans in attendance.

HFA: Home Field Advantage.

Lombardi trophy: The trophy award to the winner of the Super Bowl.

MLB: Major League Baseball.

NBA: National Basketball Association.

NCAAF: National Collegiate Athletic Association Football.

NFC: National Football Conference. One of two conferences in the NFL consisting of 16 teams.

NFL: National Football League. Consisting of two conferences - AFC and NFC.

On the road: Playing Away from home.

Punt: A kick conducted by the offense to the other team to change the field position.

PbP: Play by Play.

Point Spread: A way to create a near-even handicap between two teams by giving the inferior team a numerical point advantage.

Quarterback: An offensive player who calls the signals and directs the offensive plays of the team.

Sportsbook: A gambling establishment that takes wagers on sporting events and pays out winnings.

Super Bowl: Annual championship game between the AFC and NFC Championship winners.

Timeout: A stoppage in the game. Each team is allocated 3 timeouts per half. Timeouts are not allowed to be carried over from half to half.

Turnover on downs: When a team's offense has used all their downs but has not progressed past the required yardage to earn another set of downs. The resulting turnover gives possession to the team currently on defense.

Wildcard: First round of playoffs.

XFL: Alternative American Football League.

4th Down: A team's last chance to move the ball a required amount of yards. Teams have three options on fourth down: to punt the ball away, to kick a field goal, or to go for it.

8.3 Code Listing

GitLab link to source code:

- https://gitlab.com/mmullin98/fyp-nfl-analysis

File	Description
(1.1)HFA.ipynb	Prelim analysis
(1.2)HFA.ipynb	Prelim analysis
(1.3)HFA.ipynb	High view on HFA
(1.4)HFA.ipynb	Favourites, spread lines, totals
(1.5)HFA.ipynb	Using Soccer HFA formula to calculate HFA
(1.6)HFA.ipynb	Home Team win % produce graph, team specific work
(1.7)HFA.ipynb	Model testing
(1.8)HFA.ipynb	Home team win % including ties
(1.9)HFA.ipynb	Map spread to win % & cleaned historical scores
(1.10)HFA.ipynb	Favourite win % vs Betting Estimate %
(1.11)HFA.ipynb	Mapping home spread to win %
(1.12)HFA.ipynb	Home win % vs Betting Estimate %
(1.13)HFA.ipynb	HFA calculation
(1.14)HFA.ipynb	HFA Referee Bias
(1.15)HFA.ipynb	Super Bowl winners HFA
(1.16)HFA.ipynb	HFA by year
(1.17)HFA.ipynb	Super Bowl winners pt2
(2.1)4th Down.ipynb	Creating 4th Down dataset
(2.2)4th Down.ipynb	Conversion rates and attempts
(2.3)4th Down.ipynb	First model attempt
(2.4)4th Down.ipynb	Cleaning
(2.5)4th Down.ipynb	Run/Pass graph creation
(2.6)4th Down.ipynb	Second model attempt and full KNN creation
(2.7)4th Down.ipynb	4th Down attempts
(2.8)4th Down.ipynb	4th1-3 analysis
(2.9)4th Down.ipynb	Cleaning
(2.10)4th Down.ipynb	GUI implementation for model
(3.1)COVID.ipynb	Creation of COVID-19 years (before, during, and after)
(3.2)COVID.ipynb	Favourite spread and total points distributions
(3.3)COVID.ipynb	HFA analysis
(3.4)COVID.ipynb	4th Down analysis

Table 8.4: Summary of Python notebooks

Bibliography

1. Jamieson, J. P. The home field advantage in athletics: A meta-analysis. *Journal of Applied Social Psychology* **40,** 1819–1848 (2010).
2. Cadena, A. Open Source Football: Analyzing Home Field Advantage in the NFL. https://www.opensourcefootball.com/posts/2021-01-11-hfa-analysis/ (Jan. 2021).
3. Burke, B. *4th Down Analysis* https://www.yummymath.com/wp-content/uploads/4th_Down.pdf.
4. Baldwin, B. *4th Down Analysis Bot* https://twitter.com/ben_bot_baldwin.
5. Baker, H. P. *et al.* The Injury Rate in National Football League Players Increased Following Cancellation of Preseason Games Because of COVID-19. *Arthroscopy, Sports Medicine, and Rehabilitation* (2021).
6. Leitner, M. C. & Richlan, F. No Fans–No Pressure: Referees in Professional Football During the COVID-19 Pandemic. *Frontiers in Sports and Active Living,* 221 (2021).
7. Colomer, C. M. E., Pyne, D. B., Mooney, M., McKune, A. & Serpell, B. G. Performance Analysis in Rugby Union: a Critical Systematic Review. en. *Sports Medicine - Open* **6,** 4. ISSN: 2199-1170, 2198-9761. https://sportsmedicine-open.springeropen.com/articles/10.1186/s40798-019-0232-x (Dec. 2020).
8. Oberstone, J. Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success. en. *Journal of Quantitative Analysis in Sports* **5.** ISSN: 1559-0410. https://www.degruyter.com/document/doi/10.2202/1559-0410.1183/html (Jan. 2009).
9. Vaz, L., Carreras, D. & Kraak, W. Analysis of the effect of alternating home and away field advantage during the Six Nations Rugby Championship. *International Journal of Performance Analysis in Sport* (Dec. 2012).
10. fantasydatapros & BenjaminDominguez. *nflfastPy* 2021. https://github.com/fantasydatapros/nflfastpy.
11. Merriam-Webster. *Merriam-Webster.com dictionary* https://www.merriam-webster.com/dictionary/garbage%5C%20time (1999).
12. Crabtree, T. *NFL Scores and betting data* 2021. https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data?select=spreadspoke_scores.csv.
13. Chang, W. & Ran, M. The Impacts of Home-Court Advantage in the NBA. en, 23.
14. Florio, M. 12th Man sets new noise record. https://profootballtalk.nbcsports.com/2013/12/02/12th-man-sets-new-noise-record/ (2013).
15. Panayotovich, S. *NFL Odds: How Much is home-field advantage really worth on the spread?* https://www.foxsports.com/stories/nfl/nfl-odds-how-much-home-field-advantage-worth-spread.
16. Moskowitz, T. J. (J. *Scorecasting : the hidden influences behind how sports are played and games are won / Tobias Moskowitz and L. Jon Wertheim.* 1st ed. ISBN: 9780307591791 (Crown Archetype, New York, 2011).
17. Neave, N. & Wolfson, S. Testosterone, territoriality, and the 'home advantage'. *Physiology Behaviour* **78,** 269–275 (2003).
18. Carre, J., Muir, C., Belanger, J. & Putnam, S. Pre-competition hormonal and psychological levels of elite hockey players: Relationship to the 'home advantage'. *Physiology & Behavior* **89,** 392–398. ISSN: 00319384. https://linkinghub.elsevier.com/retrieve/pii/S0031938406003003 (Oct. 2006).
19. *Football field goal range* https://www.rookieroad.com/football/strategy/field-goal-range/ (2022).
20. Viqtorysports. *What is a punt in American Football? Complete Guide* https://www.viqtorysports.com/what-is-a-punt-in-american-football/#:~:text=The%5C%20average%5C%20punt%5C%20in%5C%20the,yards%5C%20if%5C%20there's%5C%20no%5C%20return..

21. Statista. *Coronavirus: impact on the sports industry worldwide* tech. rep. (Statista, 2020). https://ncs4.usm.edu/pdf/covid-resources/statista-impact.pdf.
22. Seifert, K. How the NFL navigated COVID-19 this season: 959,860 tests, $100 million and zero cancellations. https://www.espn.com/nfl/story/_/id/30781978/how-nfl-navigated-covid-19-season-959860-tests-100-million-zero-cancellations (2021).
23. Buchmasser, B. *NFL releases final Covid-19 testing numbers for the 2020 season* Feb. 2021. https://www.patspulpit.com/2021/2/10/22276075/nfl-coronavirus-testing-numbers-2020-season.
24. Fourvertsfootball.com. *What Is A False Start In Football?* Dec. 2021. https://fourvertsfootball.com/what-is-a-false-start-in-football/.
25. *Defensive Pass Interference | NFL Football Operations* https://operations.nfl.com/the-rules/nfl-video-rulebook/defensive-pass-interference/.
26. *Quarterback sneak* Page Version ID: 1063732549. Jan. 2022. https://en.wikipedia.org/w/index.php?title=Quarterback_sneak&oldid=1063732549.
27. Python. *tkinter package* https://docs.python.org/3/library/tkinter.html#module-tkinter.
28. SportsDataIO. *NCAAF API Documentation* https://sportsdata.io/developers/api-documentation/ncaa-football.
29. Bet365. *Las Vegas Raiders vs Jacksonville Jaguars* https://www.bet365.com/#/AC/B12/C20627865/D19/E13948800/F19/.
30. *American Football Field Dimensions & Drawings | Dimensions.com* en. https://www.dimensions.com/element/american-football-field (2022).
31. *NFL Betting Guide – How to Bet the NFL Plus Odds and Predictions* 2019. https://www.gamblingsites.com/betting-hq/nfl/.