**One-Page Handout that Explains the Dataset and Summarizes the Analysis:**

This is an NFL play by play dataset that records data of games in 2017. Some important information includes the dates of games, the offense and defence teams, as well as each play of the game with details on when it occurred and what happened, quarters, times, teams, yard lines, formations, play types etc. The data is a comprehensive overview of the strategies and outcomes in the game.

We began by analyzing the bar plots and pie charts of some simple statistics: the frequency of each down and timeout, the frequency of each penalty and which teams received the most, as well as favored formations, pass types, rush directions, and the most frequent play types. We find a number of unsurprising conclusions, namely that 1st downs are about four times more common than fourth downs, that timeout #1 is used nearly twice more than timeout #3, and that teams in the NFL are relatively constant in the number of penalties against them.

From there, we looked at the histograms for the yards gained on pass and rush plays, the yards to go until first down, and the yard line where each play was made. We find that a surprisingly large proportion of passes are incomplete, and that most passes and rushes in the NFL do not result in significant gain or loss of yardage. We also see that a significant proportion of plays in an average NFL game occur at important yard lines, PAT attempt lines and touchback lines most commonly.

We then performed a linear regression on the starting yard line of pass or rush plays against the yards gained or lost on the play. We find from this regression that there is a relatively constant relationship between starting yard line and average number of yards gained or lost, but the data have a very interesting shape (which we explain in the broader document).

Afterwards, we created a contingency table and completed a chi-squared analysis on the defensive team and the decision to pass or rush. We find that, depending on who is on defense, teams choose to rush or pass more or less (we suggest that this is likely due to it being favorable for a team to adapt to the team that they are playing against). This is one example of how our dataset can be used to inform strategy for various teams in the NFL, particularly based on situational analysis.

To determine if the number of touchdowns scored per game in the first half of the season significantly differed as opposed to the second half, we performed a permutation test on the number of touchdowns scored per game and discovered that there was no significant difference. After comparing the chi-squared given by R to one that we wrote ourselves, we looked at the distribution of timeout usage throughout each game. We created a histogram that plotted the number of times a timeout was used versus the number of seconds from the start of the half. After plotting the probability density function over each histogram, we found when the most common time to use each timeout was and how that differed between the first half and second half of the game.

Finally, we did a logistic regression of many different variables versus the yard line, the time elapsed since the start of the game, the number of days since the start of the season. We discovered some interesting results where we expected there to be no relationship, and there were some cases where we expected a relationship but there was none. More thorough analysis is relegated to the main R script.

**One-Page Document that Discusses Ethical Issues Related to Collection of the Dataset**

There are many ethical issues that could arise when collecting data. For example, when collecting data, a researcher should ask what the data is primarily used for and the impact that this data will have. Also, the researcher must make sure that the data collected is valid. Finally, the researcher must correctly determine who should have access to the dataset. A violation of any of the above may result in an ethical issue.

In the context of sports-related data collection, there are a variety of possible uses. On one hand, the data can be used to analyze strategies, improve the sport, and develop new tactics. On the other hand, the data can also be used for statistics for sports betting; this use is perhaps more morally gray. When collecting sports-related data, researchers could be considering these types of issues related to how the data could be used.

Furthermore, researchers need to make sure that the information they provide is valid. If this NFL data was incorrect, it could lead false information and conclusions about a team. This could unfairly alter public perception of a team, or simply create false data surrounding the sport in general. Therefore, researchers are ethically obligated to make sure that the data they collect is correct and objective. Another aspect of data collection is what data to report; suppose a team scores a lot of points, but also has many fouls. Then, it may be unfair or unethical to only report the score but not the penalties. Thus, the way that data is collected could benefit some teams and disadvantage others. In other words, especially with collection of sports data, researchers should be exceptionally conscious about not creating inadvertent selection bias in forming a data set. These are some ethical issues related to the objectivity and correctness of data collection.

Finally, researchers must determine the publicity of the dataset, and who should have access. This is in general not as much of a problem for sports data, because sporting events are televised, and by nature public events. However, there are still some potential issues. For example, perhaps a game’s data should only go to participating teams rather than to all competitors. Do sports teams have right to withhold data about themselves to competitors and the general public? If so, then perhaps it is unethical to make the data widely available.

Another angle on collection is from the perspective of team administrators, and players on various teams. Team administrators may object to the collection of data, because detailed statistical analysis of a whole season of NFL plays can bring light to a particular team’s strategies. If a team’s playbook is viewed as intellectual property, one might contend that publishing a data set composed of a whole season of NFL data can be seen as a violation of that intellectual property. This is because the data set can be used by competing teams to answer questions like, “when it’s third down and they’re between the 40 and 60 yard lines, what kind of formation would the Carolina Panthers be most likely to use?”. Questions like these perhaps have no good research angle, but rather could be weaponized against particular teams. An extension of this is that players could also feel violated; their “signature moves” and such could be exposed similarly.

**One-Page Document that Discusses Ethical Issue Related to Conclusions reached from the Dataset**

When making conclusions from a dataset, there are a variety of potential ethical issues. The main three things a researcher should consider are the question they are asking, the conclusion they reach, and who they share this conclusion with. Each of these topics raises ethical concerns about arriving at conclusions from datasets.

In the context of this specific NFL dataset, one important issue is determining what question the researcher would like to answer. Some questions might have different ethical implications compared to others; for example, determining whether there is a correlation between formation and yards to go may have relatively few ethical implications, but determining the correlations between fouls for unnecessary roughness / taunting and yards to go could have more ethical implications as to the nature of the sport. Therefore, when determining what question to answer, researchers should also be aware of the ethical implications of the questions they choose.

Once they have chosen a question, they will probably find an answer and arrive at a conclusion. First of all, there is an ethical obligation that the conclusions are correct, at least to the best knowledge of the researcher given the dataset. If the tests are performed incorrectly and/or the wrong conclusions are reached, this could have negative impacts on those who view and act on the results of the test, which was discussed in the other ethics document. Even if a test is correct, there is an ethical question of what a researcher does with the conclusion. It is unethical to neglect a conclusion because it is not the expected or desired result. For example if a team would like to advertise its statistics and hires someone to analyze their game data, it may be unethical for a researcher to continually run tests repeatedly and discard them until one test finally gives a desired result. The resulting report and conclusion may not be representative of the true capabilities of the team. In particular, researchers should be very wary of conflating correlation with causation—observed characteristics of a team should not be used to make normative claims or causal analyses, absent rigorous statistical methodology and appropriate disclaimers. As such, it is important for a statistician to qualify every analysis appropriately (list assumptions, clarify scope of conclusions reached), and to avoid selection bias as best as possible. If there are two teams that have similar characteristics, for example, it can be unethical for a statistician to publish an analysis of one but not the other. These are some ethical concerns with *formulating* conclusions given some analysis of a dataset.

Finally, there is the question of who the conclusions should be given to, and what impact they will have. This is also an ethical question. For example, is it unethical to sell game data analysis to teams for profit? Does this detract from the fairness of the game and create an unlevel playing field? What level of detail should be private, and what level of detail should be public? While this is not as much of an issue in the particular dataset, there is controversy as to how biometric data should be treated. Is it a violation of privacy to publish player-specific conclusions on speed, strength, balance, reaction, mobility? In general, the targeted audience of the conclusion reached also poses some potential ethical issues.