Causal Inference (INFO 3900) - Final Report

- Group Topic: Basketball - The Effect of Three Point Attempts on Winning Games
- Team members: Daanyal Agboatwalla (daa225), Harshini Cheruvu (hkc28), Carina Lau (cl2623), Josh Green (jtg227), Kate Li (kl739)

Causal Question

Describe your causal question in a way that someone who has not taken this class would understand. Why are you interested in this question? How could answering this question allow for better decision making? Include any necessary background or context. Cite outside sources you use. Answer: Our causal question is "Does shooting more three pointers increase the likelihood of winning in the NBA?" To investigate, we are using the league-wide average number of three-point attempts over the last ten years (our data set) as a threshold. Teams shooting above this threshold are classified as high-volume three-point shooters, while those below it are considered low-volume in our study. We are interested in this question since three pointers have become a major part of the NBA in recent years, sparked by NBA superstar Stephen Curry, whose exceptional shooting has redefined the game. Fans now perceive NBA games as contests heavily influenced by three-point shooting prowess. Thus, we want to answer this question to provide valuable insights for NBA coaches on whether or not they should encourage more three point shooting to maximize their probability of winning as a team.

Describe your causal question in the language of causal inference we've learned in this course: What is the treatment? What is the outcome? What are the potential outcomes? Write these out in words and in the math notation we have used in class. Answer: Causal Question: "Does shooting more three pointers increase the likelihood of winning in the NBA?"

Treatment: Number of three point attempts (over 27.5 attempts (1) or under 27.5 attempts (0)).

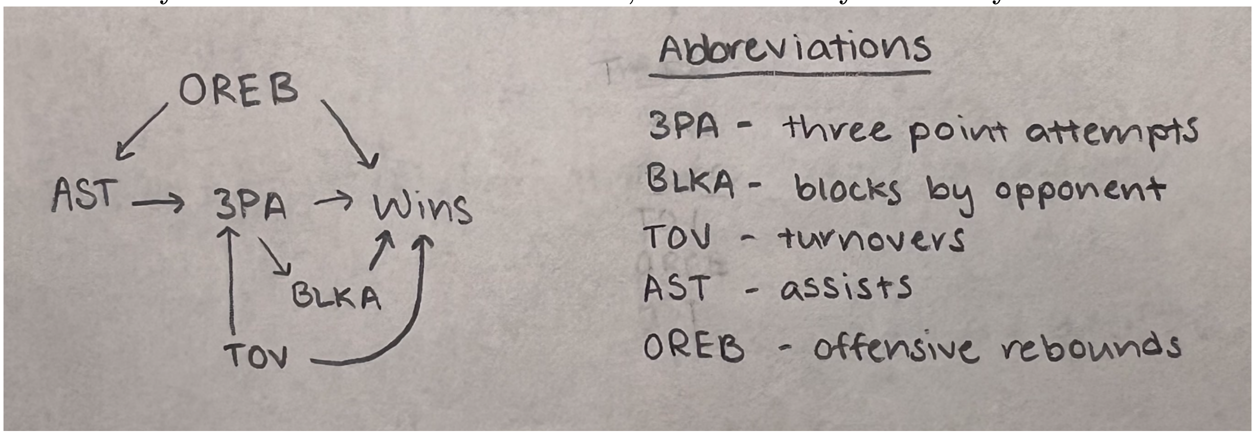
Outcome: Wins in a season (win (1) or loss (0)).

Potential Outcomes: $Y_{Team}^{Above 27.5 FG3A} = 0$, $Y_{Team}^{Above 27.5 FG3A} = 1$, $Y_{Team}^{Below 27.5 FG3A} = 0$, $Y_{Team}^{Below 27.5 FG3A} = 1$.

In words, the four potential outcomes are as follows: a team attempts over 27.5 threes and wins, a team attempts over 27.5 threes and loses, a team attempts fewer than 27.5 threes and wins, a team attempts fewer than 27.5 threes and loses.

Causal Diagram

Draw a DAG representing your causal question that includes at least three relevant variables besides treatment and outcome that are included in your dataset. You may include more than three variables. You may include variables that are not in your dataset, but at least 3 of your variables (excluding treatment and outcome) must be included in your dataset. If you use letters to denote variables, make sure they are clearly defined Answer:



3PA is the treatment variable, and wins is the outcome variable, representing whether the team wins or loses a game.

Explain your DAG: tell us in words what is meant by each edge in your DAG. Answer: OREB → Wins Offensive rebounds (OREB) provide teams with additional possessions, which create more opportunities for a team to score more points to ultimately win a game.

OREB → AST Offensive rebounds (OREB) provide teams with additional possessions, creating opportunities for more assists (AST) to be made. This edge suggests that second-chance plays often lead to more opportunities for playmaking and shot attempts.

3PA → BLKA An increased number of three point attempts (3PA) contributes to a higher possibility for the opponent to garner more blocks (BLKA).

AST → 3PA Assists (AST) facilitate ball movement, enabling players to take more efficient three-point attempts (3PA). This edge captures the teamwork and setup required for high-quality long-range shots.

3PA → Wins We hypothesize that an increased number of three-point attempts (3PA) contributes to higher potential scoring, improving the likelihood of winning (Wins). This edge assumes that volume shooting from beyond the arc is positively correlated with success.

BLKA → Wins Blocks by the opponent (BLKA) reduce a team's scoring opportunities, negatively impacting their chances of winning (Wins). This edge reflects the defensive power of the opposing team in limiting offensive efficiency.

TOV → 3PA Turnovers (TOV) can reduce offensive possessions, which may limit the opportunities for three-point attempts (3PA). This edge highlights the importance of ball control in generating shot opportunities.

TOV → Wins Turnovers (TOV) directly affect a team's ability to maintain possession and score, decreasing their chances of winning (Wins). This edge reflects the detrimental impact of lost possessions on game outcomes.

Discuss your DAG. How realistic is it? Are there variables or edges you excluded from your DAG that someone else might argue should be included? Playing devil's advocate, how would you critique the reliability of your DAG? Answer: There could definitely be certain improvements to our DAG. For instance, shooting accuracy, such as three-point shooting percentage (3P%), is a critical factor that directly affects a team's scoring efficiency and ultimately their chances of winning. Additionally, defensive metrics like steals or opponent field goal percentage are absent, Similarly, factors like player fatigue, substitutions, and game pace are excluded, even though they influence turnovers (TOV), assists (AST), and shot attempts (3PA). Certain edges are omitted for the purpose of simplification. Offensive rebounds (OREB) could reasonably have a direct link to winning (Wins), as they create second-chance scoring opportunities, bypassing the need for an intermediate relationship through three-point attempts (3PA). Another simplification that was made was deciding to not include three pointers made (3PM) as a variable. Three point attempts vs. three pointers made are variables related through instrumentality. Three point attempts acts as an intent to treat, or an instrument, while three pointers made is the treatment. We chose to simplify our analysis by not considering this relationship in our project because of the complexity of combining matching methods with instrumental variable considerations.

Taking a devil's advocate perspective, one could critique the DAG for its including only a narrow set of variables, the model risks introducing omitted variable bias, as numerous other factors undeniably influence game outcomes.

Method and Identification

What method are you using to estimate a causal effect? What causal effect are you estimating (ATE vs LATE vs ATT)? What assumptions are required to identify the causal effect via your chosen method? Answer: We are using nearest neighbor matching to estimate our causal effect with the estimand, ATT (Average Treatment effect on the Treated). In order to identify the causal effect using matching, conditional exchangeability and positivity assumptions for a given propensity score must be met.

Explain what conditional exchangeability means in the context of your causal question. Is it important? Why or why not? How do sufficient adjustment sets relate to conditional exchangeability? Answer: Conditional exchangeability is the assumption that given a set of observed covariates, the treatment assigned is independent of the potential outcomes for a unit. In the context of our causal question, this means that whether more or less than 27.5 three point shots are attempted in a game is independent of the potential outcomes (wins or losses) for the game corresponding to either treatment (above or below average 3PA). Once the sufficient adjustment set has been taken into account, there are no confounders that haven't been measured that could possibly affect the treatment and the outcome.

Conditional exchangeability is extremely important because it guarantees any difference in wins between teams that have $>$ 27.5 3PA/game and teams that have $<$ 27.5 3PA/game is because of the difference in 3PA itself and not because of other confounding factors resulting from high/low 3PA. An example could be if the overall skill level of a team has an effect on the amount of 3PA and also on their winning. Not adjusting for skill level would violate conditional exchangeability and could cause the estimate to be biased.

Sufficient adjustment sets relate to conditional exchangeability because they outline the covariates that can be controlled to ensure that the treatment assigned is independent of the potential outcomes for a given unit. In our case, AST, OREB, TOV and BLKA are all examples of covariates that have to be included in the sufficient adjustment set in order to meet the conditional exchangeability assumption. Our DAG shows that these variables are linked to both three-point attempts and the likelihood of a team winning. If we did not have the conditional exchangeability assumption, the ATT might not be as accurate because the observed relationship between our treatment and outcome could be due to confounding factors instead of true causation between our treatment, three point attempts, and outcome, winning the game.

Assuming your DAG is true, list out all non-causal paths between treatment and outcome and list one sufficient adjustment set to identify the causal effect of the treatment on the outcome. If a sufficient adjustment set does not exist, add additional variables to your DAG so that one does exist. Answer: 1. 3PA $\leftarrow$ TOV $\rightarrow$ Wins 2. 3PA $\rightarrow$ BLKA $\rightarrow$ Wins 3. 3PA $\leftarrow$ AST $\leftarrow$ OREB $\rightarrow$ Wins

A sufficient adjustment set to identify the causal effect of the treatment (three-point attempts) on the outcome (wins) includes the following variables:

- Turnovers (TOV): Controls for the impact of lost possessions on the likelihood of winning.
- Blocks Against (BLKA): Adjusts for the effect of defensive actions by the opposing team that directly impact scoring chances.
- Offensive Rebounds (OREB): Accounts for second-chance scoring opportunities that could independently influence the outcome.

Conditioning on this set blocks the non-causal paths between the treatment and the outcome, ensuring that the analysis focuses on the direct causal relationship between three-point attempts and winning games.

Discuss the plausibility of conditional exchangeability in your setting. If your sufficient adjustment set contains variables that are not in your dataset, discuss the implications. Answer: The plausibility of conditional exchangeability in our setting is dependent on whether our DAG and the dataset we used sufficiently contain all the significant confounders that have an effect on both 3PA and the outcome of NBA games (win or loss). Our DAG has variables that are important such as AST, OREB, TOV and BLKA, but could be missing other significant variables such as 3 point shooting %, player's skill levels (through advanced stats derived from player-tracking data), pace of the game (seconds per possession or FGA/game) and defensive metrics (steals, opponent's FG%, defensive rating, etc.) These unmeasured variables could add confounding and violate the conditional exchangeability assumption.

Not adjusting for these confounders could potentially add bias to the ATT and reduce the credibility of our results. An example could be that teams with a high three-point-shooting % could win more games irrespective of their 3PA (and thus, also how many 3-pointers they actually make). This would confound the relationship between 3PA and a team's wins and losses. This would also decrease the reliability of the results because there is not enough data on some of the covariates - meaning that the assumption of conditional exchangeability could fail. Even if the matching process yields a good covariate balance for the variables that are observed, there still could be unmeasured confounding variables that could lead to a situation where the treatment assignment is not actually independent of the potential outcomes. In order to fix this, we would need additional data that includes these previously omitted variables. Including additional metrics such as three-point-percentage and game speed, may increase the reliability of our analysis because this would account for more variables that affect three-point-attempts and wins.

Discuss any other identification assumptions for your method here, such as positivity and consistency. What do they mean in the context of your causal question and are they plausible? Answer: The positivity assumption requires that all teams in the dataset have a non-zero probability of attempting either above or below the threshold of 27.5 three-pointers per game, regardless of their covariates. In our study, this is largely plausible as NBA teams exhibit a range of strategies influenced by coaching decisions, player skill sets, and game contexts. However, earlier seasons, where the three-point strategy was less prevalent, may challenge this assumption, so we made sure to verify that positivity was met for all seasons included in our analysis. The consistency assumption ensures that the observed win/loss outcomes reflect the potential outcomes under the actual level of three-point attempts, assuming no other factors interfere. This is mostly reasonable in our study but could be violated by unmeasured variables such as player fatigue or the opposition's defensive strength. Addressing these factors by including additional covariates, such as three-point accuracy and defensive metrics, would improve the reliability of our results. However, adding these variables is not very plausible because it would greatly increase the number of predictors used in our model, and decrease the ease of interpretation. Together, these assumptions are critical for ensuring the reliability of our causal estimates and their relevance to understanding the impact of three-point attempts on game outcomes.

Discussion: Analysis and Results

Give some context for your dataset. Who is included in your dataset? How was the data collected? When was the data collected? Make sure to cite the dataset. Answer: Our dataset is a csv that includes all of the NBA box score stats from 2010 to 2024. This includes columns such as season year, team name, three point attempts, three point field goals, turnovers, offensive rebounds, defensive rebounds, etc. It includes many metrics that are measured about the performance of a team in a specific game. There are thousands of rows in this dataset, with each game that was played and the statistics/metrics corresponding to each team. Although the method of data collection is unknown, it can be assumed that it was web scraped from the NBA website or other sports analytics sites. We have also tried cross-referencing the data from specific games with the reputable NBA site to confirm that the data is accurate. The data was updated 5 months ago, May 2024. Here is the data: https://github.com/NocturneBear/NBA-Data-2010-2024

Discuss any choices you made regarding data cleaning and processing: Did your data have missing values or outliers? How did you handle them? Were there any variables you dichotomized (i.e. made binary), or variables that you changed the format (e.g. yes/no to 1/0)? Answer: Our data did not have any missing values. It was checked in our R code, and we saw that there were no missing values to deal with. However, we did notice that we had outliers. We decided to keep the outliers in our dataset since we considered it natural variation in our data, as teams are bound to shoot substantially more or less three pointers in specific games over others, just by chance. Moreover, we also saw that there were similar numbers of outliers for both above and below the threshold of 27.5 three point attempts, so based on these observations, we made the decision that it was reasonable to keep outliers.

Again, we made the decision to dichotomize the treatment, making it a binary of whether or not a team shot above or below the league-wide average over the last 10 years of 27.5 three point attempts. We also dichotomized the outcome, wins and losses, with a win corresponding to a value of 1 and a loss corresponding to a value of 0.

Discuss the impact of any choices you made regarding your dataset, such as choices you made in data cleaning or processing. Answer: In our data cleaning and processing, we created a for loop in our R code to change the names of the seasons. We made the names of the seasons integers so that it would be easier to graph out without issues with type. Moreover, another decision we made in our data cleaning and processing was that we spliced our dataset to only include metrics we were interested in using for our DAGs and other analyses, such as 3 point attempts, turnovers, and rebounds. Lastly, we also added in another column manually that includes binary values of whether or not a team shoots above or below 27.5 three point attempts (1 corresponds to shooting over and 0 corresponds to shooting less). This is useful in our analysis, as it essentially represents our treatment.

Explain how you estimated a causal effect.

- If you used matching, explain and discuss your choices. What formula did you use and why? What matching strategy did you use and why? Are there any advantages or drawbacks to the strategy you chose? How many units did your matching drop? How was the covariate balance in your matched sample? Discuss the implications of any choices you made and the quality of your matching.
- If you didn't use matching, explain any choices you made related to the method you used and discuss their implications. Think about advantages or drawbacks to any choices you made, possible bias-variance trade-offs, and assessing how well your method did.

Answer: We used the nearest neighbor matching strategy because it produced the least number of unmatched units and the best covariate balance as assessed by standardized mean differences (SMDs), variance ratios, and eCDF maxes. In our analysis, we originally attempted to conduct a matching analysis of the most recent seasons (2019-2024), which have increased three point attempts across the NBA compared to previous seasons. However, this method resulted in very poor covariate balance and left over 80% of the units unmatched, which made it difficult to conclude anything from that analysis. Thus, we attempted then attempted matching with nearest neighbor and coarsened exact matching (though optimal is preferred, we had too much data for it to produce timely results) on all seasons contained in the dataset after the processing step. Comparison of the covariate balance and the number of matched vs. dropped units between the nearest neighbor and the coarsened exact matching revealed that the nearest neighbor method was better suited for our data.

From the results of our analysis, 16375 units were matched, while 566 units were unmatched and thus dropped. This is a fairly good amount of matches considering the amount of data we had.

The covariate balance in our matched sample was quite good. The absolute values of the standardized mean differences (SMDs) for all the covariates were less than 0.1, the variance ratios were all very close to 1, and the eCDF max values were close to 0, indicating a good balance. These thresholds were obtained from the CRAN webpage linked in the Task 3 & 4 check-in document: https://cran.r-project.org/web/packages/MatchIt/vignettes/assessing-balance.html#assessing-balance-with-matchit

Our matching has high quality in terms of covariate balance and ratio of matched units to dropped units. However, there is an imbalance in data for seasons where the three point attempts became a popular strategy (around 5 years) vs. the years prior to this (around 10 years), so this is a potential issue with the interpretability and value of our results.

Report your causal effect estimate and interpret it in the context of your causal question.
Answer: Our causal effect estimate (ATT) is -0.0097. Since this estimate is close to 0, it seems that a high number of three point attempts by an NBA team (defined in our analysis as >27.5) does not cause the team to win. Based on this conclusion, NBA teams perhaps should not focus on three point attempts in their game strategy, as it appears to not have a convincing causal effect on the game's outcome.

Discuss the limitations of your analysis: what are the limitations of your dataset? Is there other data you would have wanted to have to bolster your analysis? Playing devil's advocate, how would you critique the reliability of you causal estimate? **Answer:** Our causal effect estimate (ATT) is -0.0097. Since this estimate is close to 0, it seems that a high number of three point attempts by an NBA team (defined in our analysis as >27.5) does not cause the team to win. Based on this conclusion, NBA teams perhaps should not focus on three point attempts in their game strategy, as it appears to not have a convincing causal effect on the game's outcome.

The NBA has only recently popularized the three point attempt in game plans, so we have limited data on this new strategy, compared to the data we have for years prior to this change. Having more data on NBA seasons after the heavy incorporation of the three point attempt in games would allow us to focus our analysis on those seasons alone. This was something we considered, but were unable to do because of the lack of sufficient control units to match with the treated units when only looking at the past three to five seasons.

Some other issues with the reliability of our causal estimate include our estimand (ATT), the limitations of our selected matching method (nearest neighbor), and the matched vs. unmatched units in our analysis.

The ATT tends to be subject to selection bias because we are only looking at the treated units, which may or may not reflect the whole population. For instance, if we had the top 10 teams making a high number of three point attempts, the effect of the three point attempts on these teams' game outcomes would potentially be confounded by the teams' existing ranking/abilities.

Another issue with the reliability of our causal estimate is that nearest neighbor matching is a "greedy" matching method. It looks for the closest match to one of the covariates in the sufficient adjustment set, so it may not find the best quality matches with respect to all of the covariates in the sufficient adjustment set. This lack of balanced or "optimal" matching may lead to bias in our estimates as well because the units that are matched may not actually be that similar.

And finally, a good route for future analysis would be to include three point attempts as an instrumental variable, and three points made as a treatment variable, which was mentioned in section 1.2 of our report. We are currently unable to incorporate the instrumental variables in our matching analysis, however this would add more complexity and potentially, more reliability to our analysis.

Code:

```r
# Formatting the document
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.3
```

```r
# Importing needed libraries
library(MatchIt)
```

```
## Warning: package 'MatchIt' was built under R version 4.3.3
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Data Cleaning and Processing

```r
# Loading and viewing the data set
season_totals <- read.csv("regular_season_totals_2010_2024.csv")
head(season_totals)
```

```
##   SEASON_YEAR    TEAM_ID TEAM_ABBREVIATION              TEAM_NAME  GAME_ID
## 1     2022-23 1610612744               GSW  Golden State Warriors 22201230
## 2     2020-21 1610612749               MIL        Milwaukee Bucks 22000051
## 3     2013-14 1610612751               BKN          Brooklyn Nets 21300359
## 4     2013-14 1610612757               POR Portland Trail Blazers 21300347
## 5     2018-19 1610612745               HOU        Houston Rockets 21801200
## 6     2012-13 1610612745               HOU        Houston Rockets 21200718
##             GAME_DATE     MATCHUP WL MIN FGM FGA FG_PCT FG3M FG3A FG3_PCT FTM
## 1 2023-04-09T00:00:00   GSW @ POR  W  48  58  96  0.604   27   49   0.551  14
## 2 2020-12-29T00:00:00   MIL @ MIA  W  48  51  92  0.554   29   51   0.569  13
## 3 2013-12-16T00:00:00 BKN vs. PHI  W  48  47  78  0.603   21   35   0.600  15
## 4 2013-12-14T00:00:00   POR @ PHI  W  48  52  93  0.559   21   37   0.568  14
## 5 2019-04-07T00:00:00 HOU vs. PHX  W  48  53 100  0.530   27   57   0.474  16
## 6 2013-02-05T00:00:00 HOU vs. GSW  W  48  46  91  0.505   23   40   0.575  25
##   FTA FT_PCT OREB DREB REB AST TOV STL BLK BLKA PF PFD PTS PLUS_MINUS GP_RANK
## 1  16  0.875    9   49  58  47  16  13   6    3 18   9 157         56       1
## 2  15  0.867   10   35  45  32  17  14   2    6 22  18 144         47       1
```

```
## 3  26  0.577    4   38  42  35  21  10   2    0  20   24  130          36        1
## 4  20  0.700   15   31  46  41  19   9   8    5  16   15  139          34        1
## 5  21  0.762   12   40  52  34   9  12   5    0  16   18  149          36        1
## 6  35  0.714   14   34  48  35   9   6   2    5  24   26  140          31        1
##   W_RANK L_RANK W_PCT_RANK MIN_RANK FGM_RANK FGA_RANK FG_PCT_RANK FG3M_RANK
## 1      1      1          1      159        3      316          35         1
## 2      1      1          1      113       57      559         122         1
## 3      1      1          1      159       69     1805          14         1
## 4      1      1          1      159       11      212          82         1
## 5      1      1          1      135       34      153         264         1
## 6      1      1          1      151       72      229         472         1
##   FG3A_RANK FG3_PCT_RANK FTM_RANK FTA_RANK FT_PCT_RANK OREB_RANK DREB_RANK
## 1        56           34     1831     2089         383      1392         8
## 2        25           16     1536     1776         414       847       927
## 3        38           35     1548      799        2337      2386       253
## 4        19           74     1687     1557        1773       319      1280
## 5         9          218     1367     1374        1309       649       366
## 6         8           73      173      107        1626       492       604
##   REB_RANK AST_RANK TOV_RANK STL_RANK BLK_RANK BLKA_RANK PF_RANK PFD_RANK
## 1       47        1     1578       55      512       457     678     2455
## 2      919      146     1636       33     1830      1361    1552     1220
## 3     1221       22     2280      406     1979         1    1004      453
## 4      664        2     2065      628      204      1269     272     2190
## 5      340       84      176      149      917         1     244     1725
## 6      365        9      112     1623     2035      1115    1974      154
##   PTS_RANK PLUS_MINUS_RANK AVAILABLE_FLAG
## 1        3               1              1
## 2       14               7              1
## 3       19              12              1
## 4        4              15              1
## 5        4              20              1
## 6        2              28              1
```

Filtering dataset

```r
# Extract columns of interest (FG3A, Year, WL, OREB, TOV, BLKA) and team name

filtered_data <- season_totals[, c("SEASON_YEAR", "TEAM_NAME", "WL", "FG3A", "TOV", "OREB", "BLKA")]

# For loop to change season names to the first year in the range
# Done to simplify our year variable for graphing and easier interpretation

for (i in 1:length(filtered_data$SEASON_YEAR)) {
  if (filtered_data$SEASON_YEAR[i] == "2023-24") {
    filtered_data$SEASON_YEAR[i] <- 2023
  } else if (filtered_data$SEASON_YEAR[i] == "2022-23") {
    filtered_data$SEASON_YEAR[i] <- 2022
  } else if (filtered_data$SEASON_YEAR[i] == "2021-22") {
    filtered_data$SEASON_YEAR[i] <- 2021
  } else if (filtered_data$SEASON_YEAR[i] == "2020-21") {
    filtered_data$SEASON_YEAR[i] <- 2020
  } else if (filtered_data$SEASON_YEAR[i] == "2019-20") {
```

```
    filtered_data$SEASON_YEAR[i] <- 2019
  } else if (filtered_data$SEASON_YEAR[i] == "2018-19") {
    filtered_data$SEASON_YEAR[i] <- 2018
  } else if (filtered_data$SEASON_YEAR[i] == "2017-18") {
    filtered_data$SEASON_YEAR[i] <- 2017
  } else if (filtered_data$SEASON_YEAR[i] == "2016-17") {
    filtered_data$SEASON_YEAR[i] <- 2016
  } else if (filtered_data$SEASON_YEAR[i] == "2015-16") {
    filtered_data$SEASON_YEAR[i] <- 2015
  } else if (filtered_data$SEASON_YEAR[i] == "2014-15") {
    filtered_data$SEASON_YEAR[i] <- 2014
  } else if (filtered_data$SEASON_YEAR[i] == "2013-14") {
    filtered_data$SEASON_YEAR[i] <- 2013
  } else if (filtered_data$SEASON_YEAR[i] == "2012-13") {
    filtered_data$SEASON_YEAR[i] <- 2012
  } else if (filtered_data$SEASON_YEAR[i] == "2011-12") {
    filtered_data$SEASON_YEAR[i] <- 2011
  } else if (filtered_data$SEASON_YEAR[i] == "2010-11") {
    filtered_data$SEASON_YEAR[i] <- 2010
  }
}

# Turning the season year variable into integer type
filtered_data$SEASON_YEAR <- as.integer(filtered_data$SEASON_YEAR)

# View filtered dataframe
head(filtered_data)
```

```
##   SEASON_YEAR              TEAM_NAME WL FG3A TOV OREB BLKA
## 1        2022  Golden State Warriors  W   49  16    9    3
## 2        2020        Milwaukee Bucks  W   51  17   10    6
## 3        2013          Brooklyn Nets  W   35  21    4    0
## 4        2013 Portland Trail Blazers  W   37  19   15    5
## 5        2018        Houston Rockets  W   57   9   12    0
## 6        2012        Houston Rockets  W   40   9   14    5
```

```
# We also find that the average of the number of 3 point attempts across all
# games in recorded in our dataset is ~27.5, so we made our baseline of
# comparison 27.5 three point attempts. So, in our potential outcomes, we
# considered number of three point attempts (treatment) as a binary variable.
# This variable is coded as follows: over 27.5 three point attempts is given a value
# of 1, and under 27.5 three point attempts is given a value of 0.


print(mean(filtered_data$FG3A))
```

```
## [1] 27.54703
```

```
# Adding a column to identify whether a game (row) has above/below 27.5 3PT attempts
# Adding a column with treatment where treat = 1 if the attempts are above average,
# and treat = 0 if attempts are below average.
```

```r
for (i in 1:length(filtered_data$FG3A)){
  if (filtered_data$FG3A[i] > 27.5) {
    filtered_data$treat[i] = 1
    filtered_data$ABOVE_AVG_FG3A[i] = TRUE
  } else {
    filtered_data$treat[i] = 0
    filtered_data$ABOVE_AVG_FG3A[i] = FALSE

  }
}

print(head(filtered_data))
```

```
##   SEASON_YEAR              TEAM_NAME WL FG3A TOV OREB BLKA treat ABOVE_AVG_FG3A
## 1        2022  Golden State Warriors  W   49  16    9    3     1           TRUE
## 2        2020        Milwaukee Bucks  W   51  17   10    6     1           TRUE
## 3        2013          Brooklyn Nets  W   35  21    4    0     1           TRUE
## 4        2013 Portland Trail Blazers  W   37  19   15    5     1           TRUE
## 5        2018        Houston Rockets  W   57   9   12    0     1           TRUE
## 6        2012        Houston Rockets  W   40   9   14    5     1           TRUE
```

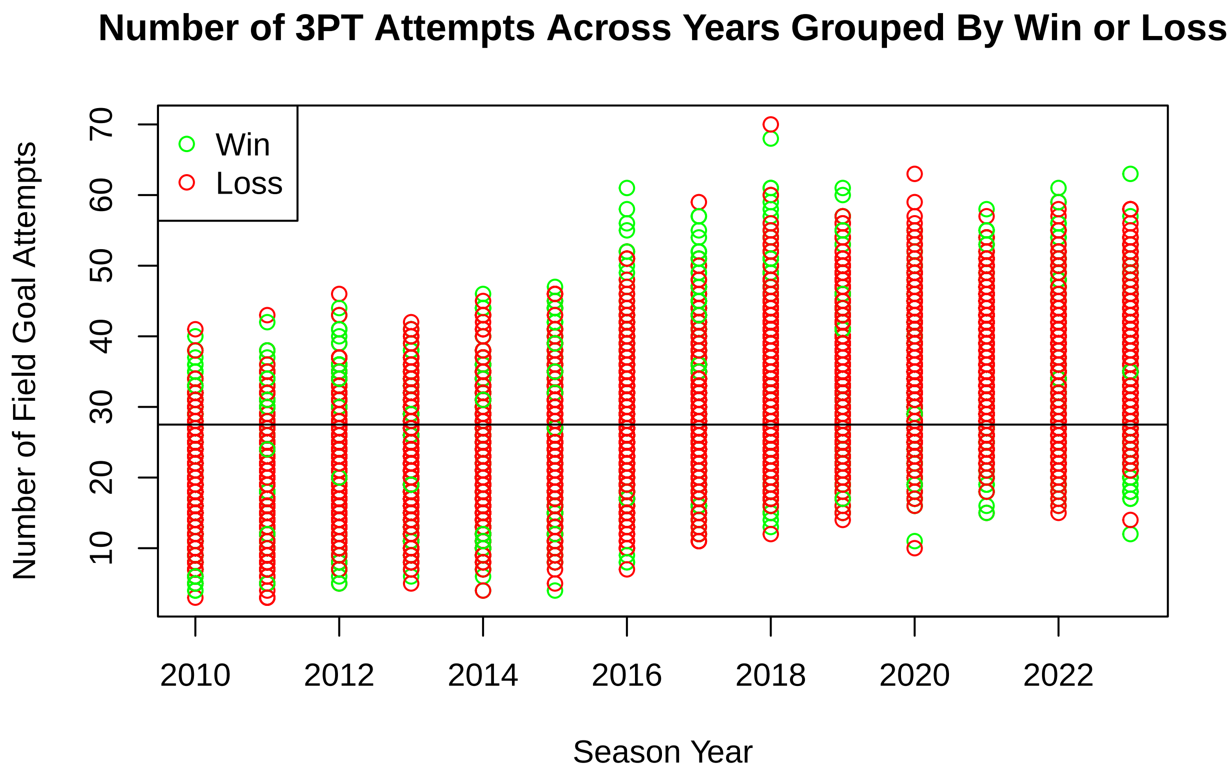
```r
# Checking for missing values
print(colSums(is.na(filtered_data)))
```

```
##    SEASON_YEAR      TEAM_NAME             WL           FG3A            TOV
##              0              0              0              0              0
##           OREB           BLKA          treat ABOVE_AVG_FG3A
##              0              0              0              0
```

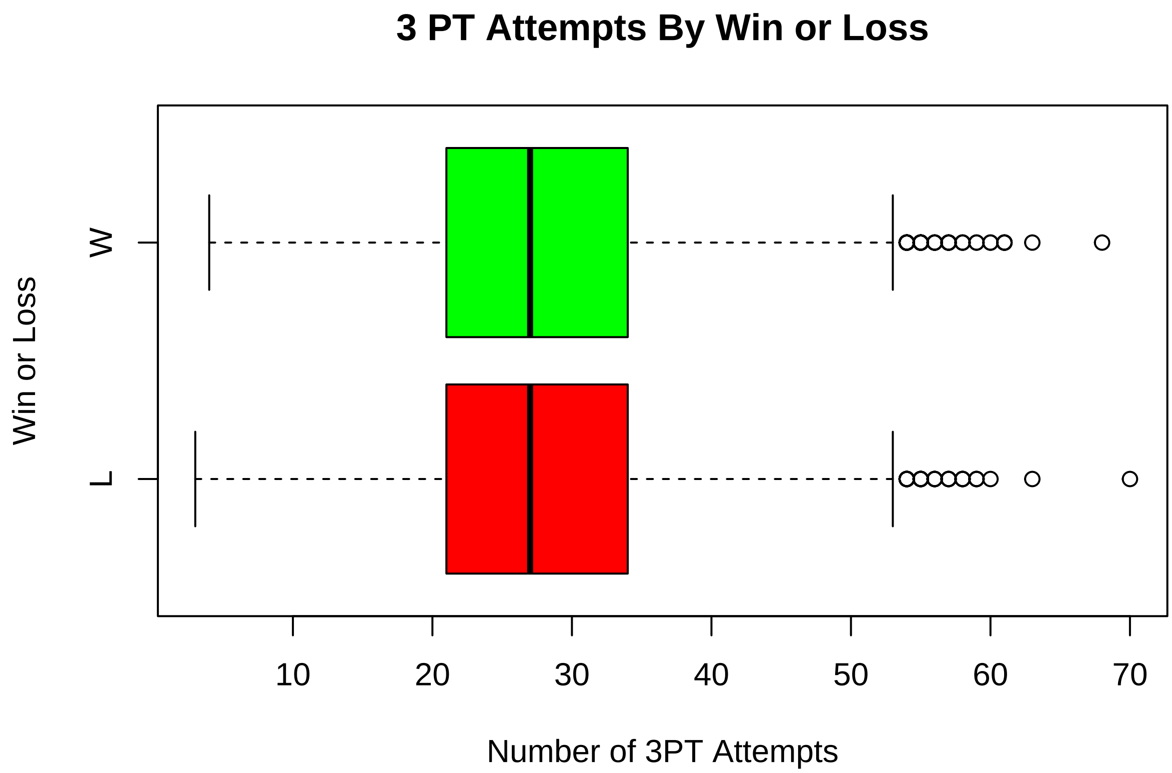
```r
# Since no NA values were found, there are no missing data
```

Data Exploration Continued (w/ Graphs)

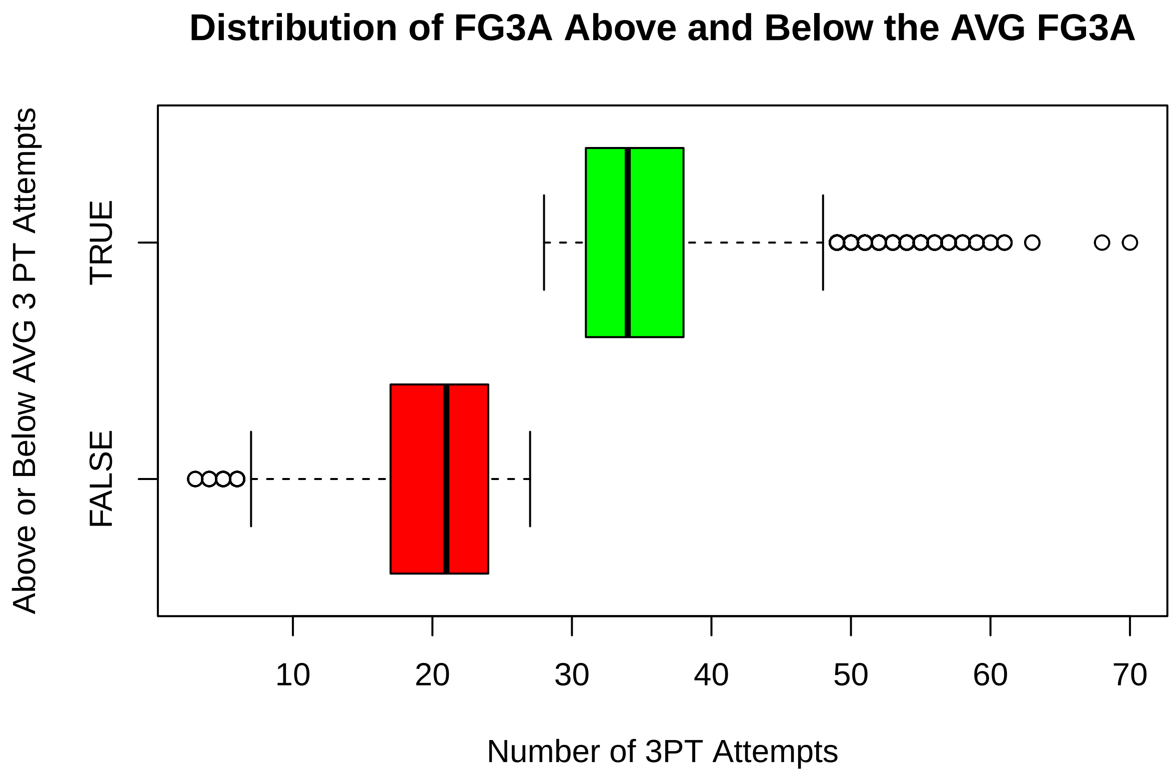
```r
# Setting wins as green and losses as red
colors <- ifelse(filtered_data$WL == "W", "green", "red")

# Plotting each season by the number of 3 point attempts
plot(filtered_data$SEASON_YEAR, filtered_data$FG3A, col = colors,
     xlab = "Season Year",
     ylab = "Number of Field Goal Attempts",
     main = "Number of 3PT Attempts Across Years Grouped By Win or Loss")

# Adding a legend to interpret the graph
legend("topleft", col = c("green", "red"), legend = c("Win", "Loss"), pch = 1)


# Setting a horizontal line at the average number of three pointers (27.5)
abline(h = 27.5)
```




```r
# Creating a box plot for all the data, one for wins and one for losses
boxplot(FG3A ~ WL, data = filtered_data, horizontal = TRUE,
        col = c("red", "green"),
        xlab = "Number of 3PT Attempts", ylab = "Win or Loss",
        main = "3 PT Attempts By Win or Loss")
```




```r
# Based on the box plots we can see that there are some outliers, but
# they exist for both the wins and the losses, which show the high volume
# three point shooting games across the years

# This is completely normal as there are certain games where NBA players will
# naturally decide to shoot more three pointers as a part of their game plan

# We decided to keep the outliers in our analysis since we believe they represent
# natural variations in our data, as NBA strategy has shifted over time

# Creating a box plot that can look at distributions of data above and
# below the avg 3PA value
boxplot(FG3A ~ ABOVE_AVG_FG3A, data = filtered_data, horizontal = TRUE,
        col = c("red", "green"),
        xlab = "Number of 3PT Attempts", ylab = "Above or Below AVG 3 PT Attempts",
        main = "Distribution of FG3A Above and Below the AVG FG3A")
```



Matching analysis

```
# Turning BLKA, TOV, OREB into integers
filtered_data$BLKA <- as.integer(filtered_data$BLKA)
filtered_data$TOV <- as.integer(filtered_data$TOV)
filtered_data$OREB <- as.integer(filtered_data$OREB)

# Formula = FG3A ~ TOV + OREB + BLKA
# where the variables in the formula are part of the sufficient adjustment set
match.output <- matchit(treat ~ TOV + OREB + BLKA,
                 data = filtered_data,
                 method = "nearest",
                 distance = "glm",
                 replace = F,
                 ratio = 1,
                 estimand = "ATT")

# Results of the matching output, shows how many units were matched
summary(match.output, interactions = F, un = F)
```

```
##
## Call:
## matchit(formula = treat ~ TOV + OREB + BLKA, data = filtered_data,
##     method = "nearest", distance = "glm", estimand = "ATT", replace = F,
```

```
##     ratio = 1)
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.4941        0.4912          0.0815     1.0665    0.0167
## TOV            13.9932       14.2461         -0.0648     1.0572    0.0084
## OREB           10.4336       10.5033         -0.0183     0.9915    0.0024
## BLKA            4.7565        4.8796         -0.0498     0.9785    0.0059
##          eCDF Max Std. Pair Dist.
## distance   0.0406          0.0816
## TOV        0.0344          0.7240
## OREB       0.0115          1.0622
## BLKA       0.0235          0.8102
##
## Sample Sizes:
##           Control Treated
## All         16941   16375
## Matched     16375   16375
## Unmatched     566       0
## Discarded       0       0
```

```r
# Saving match output data
dat <- match.data(match.output)
head(dat)
```

```
##   SEASON_YEAR              TEAM_NAME WL FG3A TOV OREB BLKA treat ABOVE_AVG_FG3A
## 1        2022  Golden State Warriors  W   49  16    9    3     1           TRUE
## 2        2020        Milwaukee Bucks  W   51  17   10    6     1           TRUE
## 3        2013          Brooklyn Nets  W   35  21    4    0     1           TRUE
## 4        2013 Portland Trail Blazers  W   37  19   15    5     1           TRUE
## 5        2018        Houston Rockets  W   57   9   12    0     1           TRUE
## 6        2012        Houston Rockets  W   40   9   14    5     1           TRUE
##    distance weights subclass
## 1 0.4975615       1        1
## 2 0.4629017       1        2
## 3 0.4949190       1        3
## 4 0.4550067       1        4
## 5 0.5692129       1        5
## 6 0.5227087       1        6
```

```r
# Changing the wins and losses to 1 and 0 (numeric values) to allow for linear
# regression
dat_numeric <- dat

# Wins take on a value of 1 and losses take on a value of 0
for (i in (1:length(dat$WL))) {
  if (dat$WL[i] == 'W') {
    dat_numeric$WL[i] = 1
  }

  else if (dat$WL[i] == 'L') {
    dat_numeric$WL[i] = 0
  }
```

```r
}

# Fitting a linear regression for the outcome (wins/losses) on the treatment
# and including the variables in the sufficient adjustment set
fit <- lm(WL ~ treat + TOV + OREB + BLKA,
          data = dat_numeric,
          weights = weights)

summary(fit)
```

```
##
## Call:
## lm(formula = WL ~ treat + TOV + OREB + BLKA, data = dat_numeric,
##     weights = weights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7477 -0.4909  0.2930  0.4753  0.8766
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.8221655  0.0135379  60.731   <2e-16 ***
## treat       -0.0096568  0.0054369  -1.776   0.0757 .
## TOV         -0.0121037  0.0007056 -17.155   <2e-16 ***
## OREB         0.0012808  0.0007319   1.750   0.0801 .
## BLKA        -0.0324858  0.0011255 -28.864   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4915 on 32745 degrees of freedom
## Multiple R-squared:  0.03368,    Adjusted R-squared:  0.03356
## F-statistic: 285.3 on 4 and 32745 DF,  p-value: < 2.2e-16
```

```r
# Estimate of causal effect: ATT (Average effect of the Treatment on the treated)
fit$coefficients[2]
```

```
##        treat
## -0.009656841
```