Sports Betting Analysis of NBA Matches Since 2007

Peter Grantcharov, Fernando Troeman, Po-Chieh Liu

12/10/2018

1 Introduction

1.1 Overview

In this report, we will investigate a dataset from Sports Book Reviews Online (https://www.sportsbookreviewsonline.com/scoresoddsarchives/nba/nbaoddsarchives.htm). This website has scraped sports betting data from online sportsbooks (websites that offer sports betting services) for all NBA basketball seasons since the 2007-2008 season. Their databases also contain the game outcomes, thereby allowing for insightful comparisons between the betting odds established before the individual games and actual game results.

1.2 Motivation

We chose this topic to satisfy the primal parts of our psyches that are stimulated by:

1) Sports
2) Money
3) Gambling

... and of course:

4) Data exploration and visualization

We therefore found this dataset to be an excellent suitor for delivering captivating insights and exercising our data analysis and visualization skills. Further, through this report's analysis, we are particularly interested in answering the questions of:

- *How good are the market and the sportsbooks at correctly determining the odds for a given match-up?*
- *Are there certain betting systems that would have been able to turn a profit if implemented throughout the 11 season span of this dataset?*

1.3 Group

Every group member was very curious and committed to answering these questions. Below is an outline of how the responsibilities were divided amongst our group:

Peter Grantcharov:

- Wrote sections 1, 2, and 3; compiled reports

Po-Chieh Liu:

- Created over/under analysis; created individual teams analysis

Fernando Troeman:

- Constructed entire interactive visualization; created spread analysis

2 Data Description

2.1 Data Collection

The data files were downloaded as individual seasons in separate *.xlsx* files from https://www.sportsbookreviewsonline.com/. Sports betting data is a highly valued commodity, so comprehensive and clean open source databases are hard to come by.

The individual seasons were then stitched together into a single *.csv* file. The script to perform this merger was written in Python, and can be found in our GitHub repository entitled *csv_merger.py*. It should also be noted that during the process of merging the data, the *Date* column was also transformed to be a date object from the Python package *Datetime* (*Datetime* formated values in CSV files are automatically read as *Date* objects in R). Since the Date object in Python under the *Datetime* package does not consider February 29 as a valid date, this exception was also handled in this merger script.

2.2 Dataset Features

2.2.1 Definitions

In the untidy dataset, each row corresponded to a team. Therefore, each NBA match would have data spanning two consecutive rows. The following columns were included in the raw, untidy dataset:

1) Date - Given in the integer format "MMDD"
2) VH - Indicator of whether team in this row was the visiting team or home team (alternates down the data frame)
3) Team - Team name
4) 1st, 2nd, 3rd, 4th - The amount of points scored by a given team in the 1st, 2nd, 3rd, and 4th quarters
5) Final - Final score for the full game **not necessarily sum of quarters if the game went to overtime*
6) Open - Contains two piece of data; both are finalized at the moment that the sportsbook opens betting for a given game
    - The over/under value for the total amount of points in the games (generally around 200)
    - The expected win margin for the team that is favored to win (hereafter denoted as spread)
7) Close - Contains the same pieces of data as the "Open" column, but are now representing their respective values when the sportsbook closes betting for a given game
8) ML - The "moneyline"; if negative (favorite), represents how much money a bettor has to bet to win $100 if the team in the row wins; if positive (underdog), represents how much money a better wins on a $100 bet
9) 2H - Contains the same data as "Open" and "Close", but now only applies to the over/under scores and the win margins (spreads) for just the 2nd half of the game; this value is finalized at the points that sportsbooks close betting for a given

2.2.2 Explanations (optional)

To avoid confusion in later sections, we have included a comprehensive section that further defines some key terms pertaining to sports betting:

1) Over/Under

The over/under is a value representing the HYPOTHESIZED total amount of points in a given time period. For this dataset, this is either for the full game or for the second half of the game.

Sports bettors can pick the "over" if they believe the number of points scored in the game will be greater than the over/under value given by the sportsbook, or they can pick the "under" if they believe that the number of points scored will be less. This is considered a 50%-50% bet, however, since sportsbooks charge a commission, the payouts for a correct over/under bet will be slightly less than simply doubling your original bet.

So for example, a sportsbook may have an over/under value of 200 points for a given game. If the final score of the game was 101-105, there would've been a total of 206 points scored, and hence, a winning bettor would have had to select the "over" bet in order to get a payout.

2) Spread

For a given game, each team will have a spread value. The spread values for opposing teams will be the negatives of each other, where the team that is the favorite will have a negative spread value (e.g. -5), and the underdog will have a positive spread value (e.g. +5). To win a spread bet, a team must win by MORE than the spread value if they are the favorite, or must NOT lose by MORE than the spread value if they are the underdog.

So for example, if the Boston Celtics are favored against the New York Knicks with a spread of -5 for Boston and +5 for New York, and a bettor believes that New York will lose by less than 5 points, then they would select the +5 spread in favor of New York. Then this bettor would win the bet if Boston's point total minus New York's point total is less than 5.

Spread bets, like over/under bets, are theoretically also 50%-50% bets, with a small commission given up for a winning bet.

3) Moneyline

Moneyline bets are easier to digest, because they solely involve selecting the winner of a match. Because of this, the payouts vary significantly from match to match based on the teams (i.e., it is no longer a 50%-50% bet). Like spread betting, the favorite will have a negative moneyline value and the underdog will have a positive moneyline value. For favorites, this value is less than or equal to -100, and corresponds to the amount of money a bettor would have to bet in order to win $100 on a correct bet. For underdogs, the value is greater than or equal to +100, and corresponds to the amount of money a bettor would win on a $100 bet.

So for example, if the moneyline for an LA Lakers vs. Houston Rockets game is -150 for LA and +140 for Houston, this means that on a correct $150 bet on LA to win the game (by any margin), you would win $100, while on a correct $100 bet on Houston, you would win $140.

3 Analysis of Quality

3.1 Tidying

The data, once merged for the past 11 seasons, had to be converted into a tidy format such that each individual NBA match had its own row, rather than two separate rows for oppoosing teams.

Before tidying, the data frame looked like this:

```
options(dplyr.width = Inf)
kable(combined[1:4,], caption = "Untidy Data")
```

Table 1: Untidy Data

Date	Season	VH	Team	1st	2nd	3rd	4th	Final	Open	Close	ML	2H
2007-10-30	2007	V	Portland	26	23	28	20	97	184	189.5	900	95
2007-10-30	2007	H	SanAntonio	29	30	22	25	106	12.5	13	-1400	5
2007-10-30	2007	V	Utah	28	34	24	31	117	214.5	212	100	105.5
2007-10-30	2007	H	GoldenState	30	21	21	24	96	3	1	-120	3

After tidying, the data frame looking like this (only columns relevant to visiting team showed for convenience):

```
kable(tidy[1:5, c(2, 4, 6, 7, 8, 9, 14, 16, 18, 22, 24)], caption = "Tidy Data")
```

Table 2: Tidy Data

Date	V	V1	V2	V3	V4	VF	OUOpen	VSpreadOpen	VMoney	OU2H
2007-10-30	Portland	26	23	28	20	97	184.0	12.5	900	95.0
2007-10-30	Utah	28	34	24	31	117	214.5	3.0	100	105.5
2007-10-30	Houston	16	27	27	25	95	191.0	-2.5	-230	99.0
2007-10-31	Philadelphia	22	28	17	30	97	190.0	6.5	255	96.5
2007-10-31	Washington	23	22	25	33	110	200.0	-1.5	-125	105.0

This process was quite extensive due to the format of the data. For example, second half betting lines in column "2H" (shown above in the untidy dataframe) contained two columns worth of data:

1) Over/Under scores;
2) Spreads

Hence, to give each of these features their own columns, an algorithm had to be constructed that determined whether a given entry in "2H" represented Over/Under or Spread data, and then filled out the tidy columns accordingly. This data frame conversion can be found below the "*# MAKE TIDY DATA FRAME*" comment in the python script entitled: *data_cleaner.py*.

3.2 Cleaning

Next, we had to clean the data. The was primarily done to handle the missing values, incorrect entries, and improper data formatting. This process was done entirely in Python, and the full script can be found in *data_cleaner.py* on our GitHub repository. Relevant excerpts of such corrections are included below.

3.2.1 Team Names:

By getting a list of the unique team names, we could see a few mistakes that needed correction:

```
unique(combined$Team)
```

```
##  [1] "Portland"      "SanAntonio"    "Utah"          "GoldenState"
##  [5] "Houston"       "LALakers"      "Philadelphia"  "Toronto"
##  [9] "Washington"    "Indiana"       "Milwaukee"     "Orlando"
## [13] "Chicago"       "NewJersey"     "Dallas"        "Cleveland"
## [17] "Memphis"       "Sacramento"    "NewOrleans"    "Seattle"
## [21] "Denver"        "Detroit"       "Miami"         "Phoenix"
## [25] "Charlotte"     "Atlanta"       "NewYork"       "Boston"
## [29] "Minnesota"     "LAClippers"    "OklahomaCity"  "Brooklyn"
## [33] "Oklahoma City" "LA Clippers"
```

Clearly, inconsistencies in spelling (Oklahoma City and LA Clippers) needed to be corrected. Additionally, the Brooklyn Nets were formerly known as the New Jersey Nets before an ownership change resulted in their minor relocation. To allow for continuation analyses of the same franchise, all "NewJersey" entries were renamed as "Brooklyn".

```
# Make spelling consistent; replace NewJersey --> Brooklyn
df = df.replace(to_replace="NewJersey", value="Brooklyn")
df = df.replace(to_replace="Oklahoma City", value="OklahomaCity")
df = df.replace(to_replace="LA Clippers", value="LAClippers")
```

3.2.2 Pick 'em:

A convention in sports betting is to name 50/50 bets as "Pick 'em" bets. In this dataset, those bets were denoted as "pk" or "PK". Further, on occasion, sportsbooks will close the book on certain games for a variety of reasons. Such games are denoted as "no line" or "NL" in this dataset. The former were given values of '0' in our data frame and the latter were given 'NA' values so those games could be easily dropped. Lastly, a few mis-entries were also identified below because they either had characters in a numerical column, or had values differing by a factor of 10. The correction process is shown below:

```
kable((combined %>% filter(`2H` == 'pk'))[1:3,],
      caption = "Pick 'em Examples *Notice column '2H'")
```

Table 3: Pick 'em Examples *Notice column '2H'

Date	Season	VH	Team	1st	2nd	3rd	4th	Final	Open	Close	ML	2H
2007-11-01	2007	V	Detroit	26	22	18	25	91	4	4	-190	pk
2007-11-03	2007	V	Portland	12	19	15	34	80	203.5	203	700	pk
2007-11-03	2007	V	Sacramento	25	31	27	19	102	189.5	189	NL	pk

```
kable(combined[c(1975,23521),], caption = "Faulty Over/Under Cases")
```

Table 4: Faulty Over/Under Cases

Date	Season	VH	Team	1st	2nd	3rd	4th	Final	Open	Close	ML	2H
2008-03-16	2007	V	LALakers	23	21	22	26	92	197.5u10	196	150	3
2016-11-17	2016	V	Chicago	25	16	25	19	85	1955.5	192.5	135	96.5

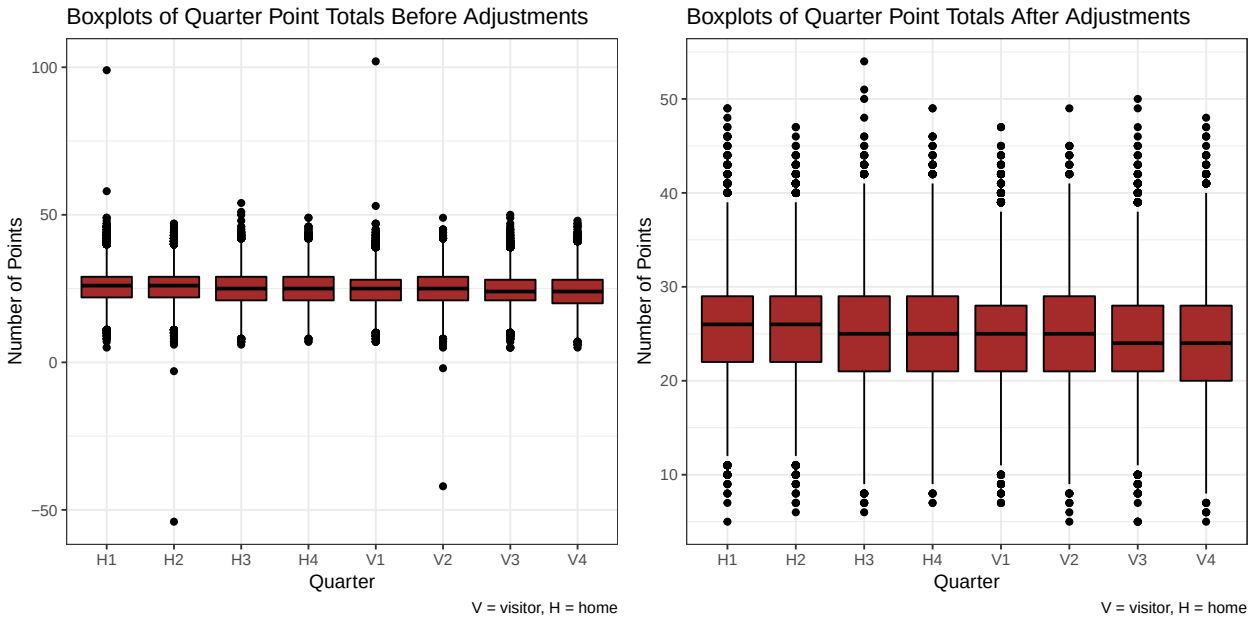
```
# Replace all pk odds (i.e. 50/50 outcomes) with 0; remove NL bets
df = df.replace(to_replace=["pk", "PK"], value=0)
df = df.replace(to_replace="NL", value=np.nan)

df.loc[df.index[1975], "Open"] = 197.5
df.loc[df.index[23521], 'Open'] = 195.5
```

3.2.3 Bad Entries:

To locate the bad entries, we found it effective to make graphs of the different variables - particularly box plots. This was performed for each of the numerical variables in the dataset:

Since all eight columns for quarter scores have nearly identical distributions, we could kill eight birds with one stone and plot them all at once. By identifying the mis-entries in these columns, we applied a generic threshold for all columns to remove the identified outliers. The before and after plots are shown below, with the Python code used to apply the corrections being printed below the graphs.

```
quarter_scores <- intermediate %>%
  select(V1, V2, V3, V4, H1, H2, H3, H4)
quarter_scores <-gather(quarter_scores, key = "Quarter", value = "Score")

before <- ggplot(quarter_scores, aes(x = Quarter, y = Score)) +
  geom_boxplot(fill = "brown", color = "black") +
  ggtitle("Boxplots of Quarter Point Totals Before Adjustments") +
  labs(x = "Quarter",
  y = "Number of Points",
  caption = "V = visitor, H = home") +
  theme_bw()

quarter_scores <- tidy %>%
  select(V1, V2, V3, V4, H1, H2, H3, H4)
quarter_scores <- gather(quarter_scores, key = "Quarter", value = "Score")

after <- ggplot(quarter_scores, aes(x = Quarter, y = Score)) +
  geom_boxplot(fill = "brown", color = "black") +
  ggtitle("Boxplots of Quarter Point Totals After Adjustments") +
  labs(x = "Quarter",
  y = "Number of Points",
  caption = "V = visitor, H = home") +
  theme_bw()

grid.arrange(before, after, ncol = 2, widths = c(9, 9))
```



Negative values are obviously impossible in the context of basketball scores, so those entries were promptly removed. Similarly, other extreme outliers were dropped, while the remaining outliers were assessed individually. Given that our dataset had over 14000 games, and that we were operating under the assumption that such mis-entries occur at random, we did not think that it would be harmful to the integrity of our analysis if we were rather loose with removing data entries. As can be seen, the quarter distributions looked much better after applying the corrections.

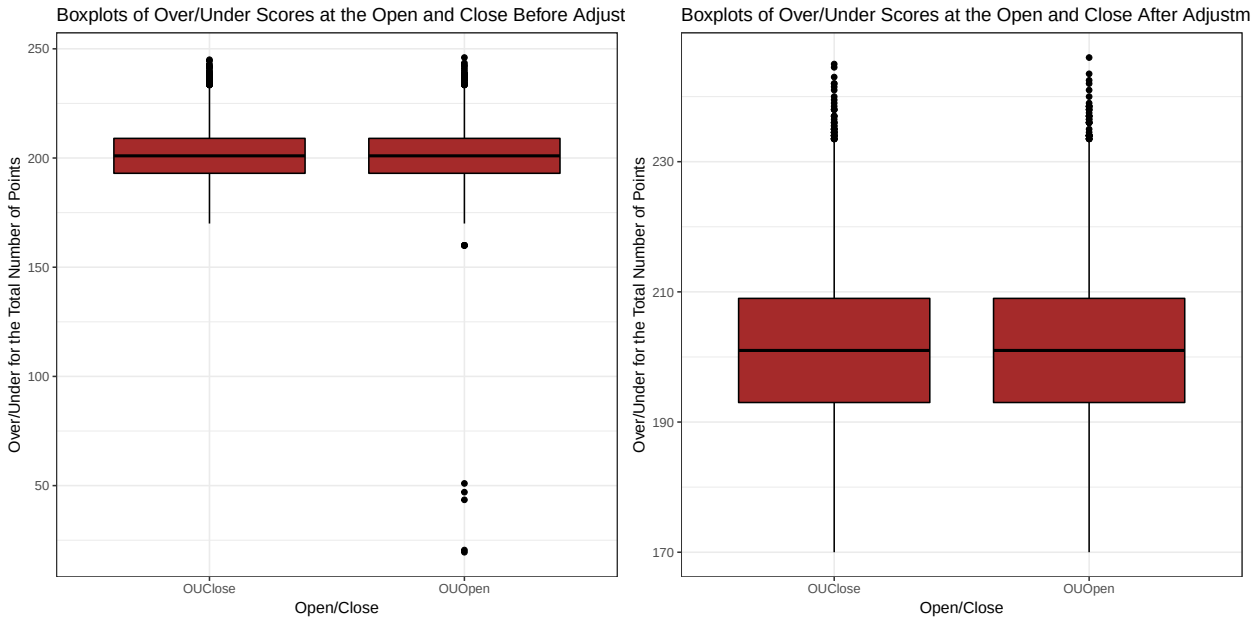
Excerpt from the cleaning script in Python:

```
# Fix outliers for quarter scores
tidy.iloc[:, 4:12] = tidy.iloc[:, 4:12][tidy.iloc[:, 4:12] < 70]
tidy.iloc[:, 4:12] = tidy.iloc[:, 4:12][tidy.iloc[:, 4:12] > 0]
```

The same process was performed for the Over/Under scores. The pre- and post-cleaning box plots appeared as follows:

```
OUs <- intermediate %>%
  select(OUOpen, OUClose)
OUs <- gather(OUs, key = "OpenClose", value = "Value")

before <- ggplot(OUs, aes(x = OpenClose, y = Value)) +
  geom_boxplot(fill = "brown", color = "black") +
  ggtitle("Boxplots of Over/Under Scores at the Open and Close Before Adjustments") +
  labs(x = "Open/Close",
  y = "Over/Under for the Total Number of Points") +
  theme_bw()

OUs <- tidy %>%
  select(OUOpen, OUClose)
OUs <- gather(OUs, key = "OpenClose", value = "Value")

after <- ggplot(OUs, aes(x = OpenClose, y = Value)) +
  geom_boxplot(fill = "brown", color = "black") +
  ggtitle("Boxplots of Over/Under Scores at the Open and Close After Adjustments") +
```

```
  labs(x = "Open/Close",
  y = "Over/Under for the Total Number of Points") +
  theme_bw()

grid.arrange(before, after, ncol = 2, widths = c(9, 9))
```



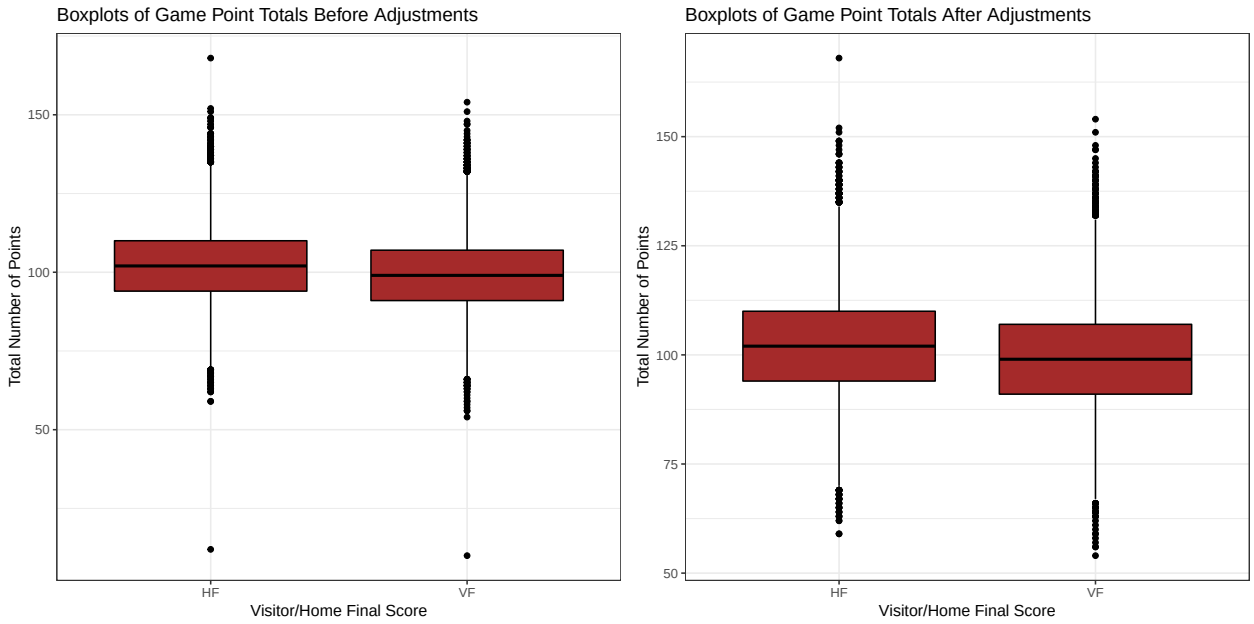
The few outliers are very obvious, so they were easily cleaned by:

```
# Fix outliers for Over/Under scores
tidy.OUOpen = tidy.OUOpen[tidy.OUOpen > 100]
```

Lastly, we reviewed the distributions of the full game scores to try to identify some mis-entries among those. Again, we found the box plot to be an effective tool for doing so:

```
Tots <- intermediate %>%
  select(VF, HF)
Tots <- gather(Tots, key = "Team", value = "Value")

before <- ggplot(Tots, aes(x = Team, y = Value)) +
  geom_boxplot(fill = "brown", color = "black") +
  ggtitle("Boxplots of Game Point Totals Before Adjustments") +
  labs(x = "Visitor/Home Final Score",
  y = "Total Number of Points") +
  theme_bw()

Tots <- tidy %>%
  select(VF, HF)
Tots <- gather(Tots, key = "Team", value = "Value")

after <- ggplot(Tots, aes(x = Team, y = Value)) +
  geom_boxplot(fill = "brown", color = "black") +
  ggtitle("Boxplots of Game Point Totals After Adjustments") +
  labs(x = "Visitor/Home Final Score",
  y = "Total Number of Points") +
```

```
  theme_bw()

grid.arrange(before, after, ncol = 2, widths = c(9, 9))
```



The outliers that appear at the bottom end are clearly also out of place, and were confirmed to be mis-entries. However, the quarter scores were correctly entered, so the VF and HF were therefore changed to be the sum of the four quarter scores. The outlier seen at the top of the "HF" boxplot was confirmed to be a legitimate score (Phoenix Suns vs. Golden State Warriors on March 15, 2009: https://www.basketball-reference.com/boxscores/200903150GSW.html), so it was not removed.

After performing these steps, and removing rows with missing data, we were pleased enough with our tidy dataset to commence our data exploration!

4 Main Analysis

As mentioned, we are particularly interested in answerings the following two questions:

- *How good are the market and the sportsbooks at correctly determining the odds for a given match-up?*
- *Are there certain betting systems that would have been able to turn a profit if implemented throughout the 11 season span of this dataset?*

To do this, we decided to explore three different avenues that we suspected might possibly reveal valuable insights:

1) Over/Under Analysis
2) Spread Analysis
3) Individual Team Analysis

4.1 Over/Under Analysis

As described in section 2.2, the over/under is a particular betting option where the gambler attempts to correctly predict whether the total amount of points in a game (combined for both teams) will be greater or less than some *value.* This *value* is selected based on various predictive models by the sportsbooks, and further, will change over the course of the betting period based on which side of the over/under the majority of gamblers are placing their money. Because of this fact, we can frequently observe changes in the over/under totals between the start of the betting session (OUOpen in our dataset) and the conclusion of the betting session (OUClose).

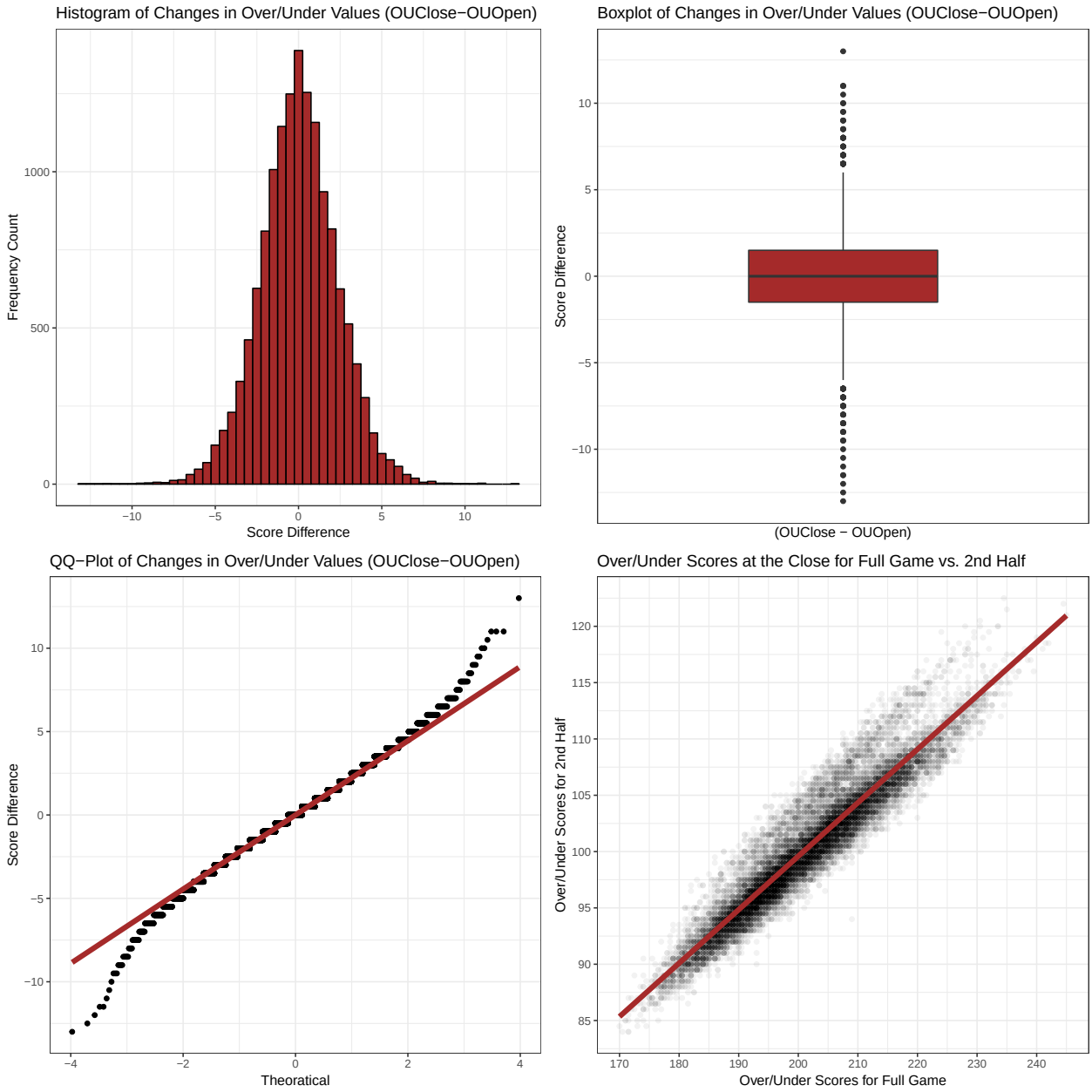
Given our intuition that changes in the over/under values carry some information (perhaps an indicator of new information regarding the game), we wanted to explore whether such swings had any relationship to the true outcomes. Additionally, we wanted to see the relatinship between the full game over/under scores and the over/under scores for the second half (OU2H in dataset). This is what we explored and tested in this section.

4.1.1 Data Overview

As a starting point, we wanted to explore the distribution of the relevant variables. As such, we plotted a histogram, box plot, and quartile-quartile plot of the differences between OUOpen and OUClose. As a means of comparison, we also plotted a scatter plot that aimed to showcase the strength of the relationship between the full game over/under values, and those for just the second half.

```
tidy <- tidy %>%
  select(Date, SZN, V, H, V1, V2, V3, V4, H1, H2, H3, H4,
         VF, HF, OUOpen, OUClose, OU2H, VMoney, HMoney) %>%
  mutate(Total = VF+HF) %>%
  mutate(Total_2H = V3+V4+H3+H4)

df <- tidy %>%
  select(OUOpen, OUClose) %>%
  mutate(Diff = OUClose - OUOpen)

histo <- ggplot(df, aes(x = Diff)) +
  geom_histogram(binwidth = 0.5, fill='brown', color='black') +
  xlab("Score Difference") +
  ylab("Frequency Count") +
  ggtitle("Histogram of Changes in Over/Under Values (OUClose-OUOpen)") +
```

```
  theme_bw()

box <- ggplot(df, aes(y= Diff)) +
  geom_boxplot(fill = "brown") +
  scale_x_discrete() +
  xlab("(OUClose - OUOpen)") +
  ylab("Score Difference") +
  ggtitle("Boxplot of Changes in Over/Under Values (OUClose-OUOpen)") +
  theme_bw()

qq <- ggplot(df, aes(sample = Diff)) +
  geom_qq() +
  stat_qq_line(distribution = qnorm, color= "brown", size = 2) +
  xlab("Theoratical") +
  ylab("Score Difference") +
  ggtitle("QQ-Plot of Changes in Over/Under Values (OUClose-OUOpen)") +
  theme_bw()

df <- tidy %>% select(OUClose, OU2H) %>% mutate(Ratio = OUClose / OU2H)

scat <- ggplot(df, aes(x = OUClose, y = OU2H)) +
  geom_point(alpha = 0.05) +
  geom_smooth(method = lm, se = FALSE, color = "brown", show.legend = TRUE, size = 2) +
  scale_x_continuous("Over/Under Scores for Full Game",
                     breaks = c(170, 180, 190, 200, 210, 220, 230, 240),
                     labels = c("170", "180", "190", "200", "210", "220", "230", "240")) +
  scale_y_continuous("Over/Under Scores for 2nd Half",
                     breaks = c(85,90,95,100,105,110,115,120),
                     labels = c("85","90","95","100","105","110","115","120")) +
  ggtitle("Over/Under Scores at the Close for Full Game vs. 2nd Half") +
  theme_bw()

grid.arrange(histo, box, qq, scat, ncol = 2)
```



The three plots pertaining to the distribution appear to be fairly standard with few clues to oddities that may be exploitable for profitable betting systems. The plot in the bottom right, however, shows us that although the second half over/unders vs. full time over/unders have a very strong trend line with a slope of about 0.5, there are several instances where these scores deviate somewhat significantly from this line, so it is by no means a "fixed" relationship. We will explore this further.

4.1.2 Preliminary Model

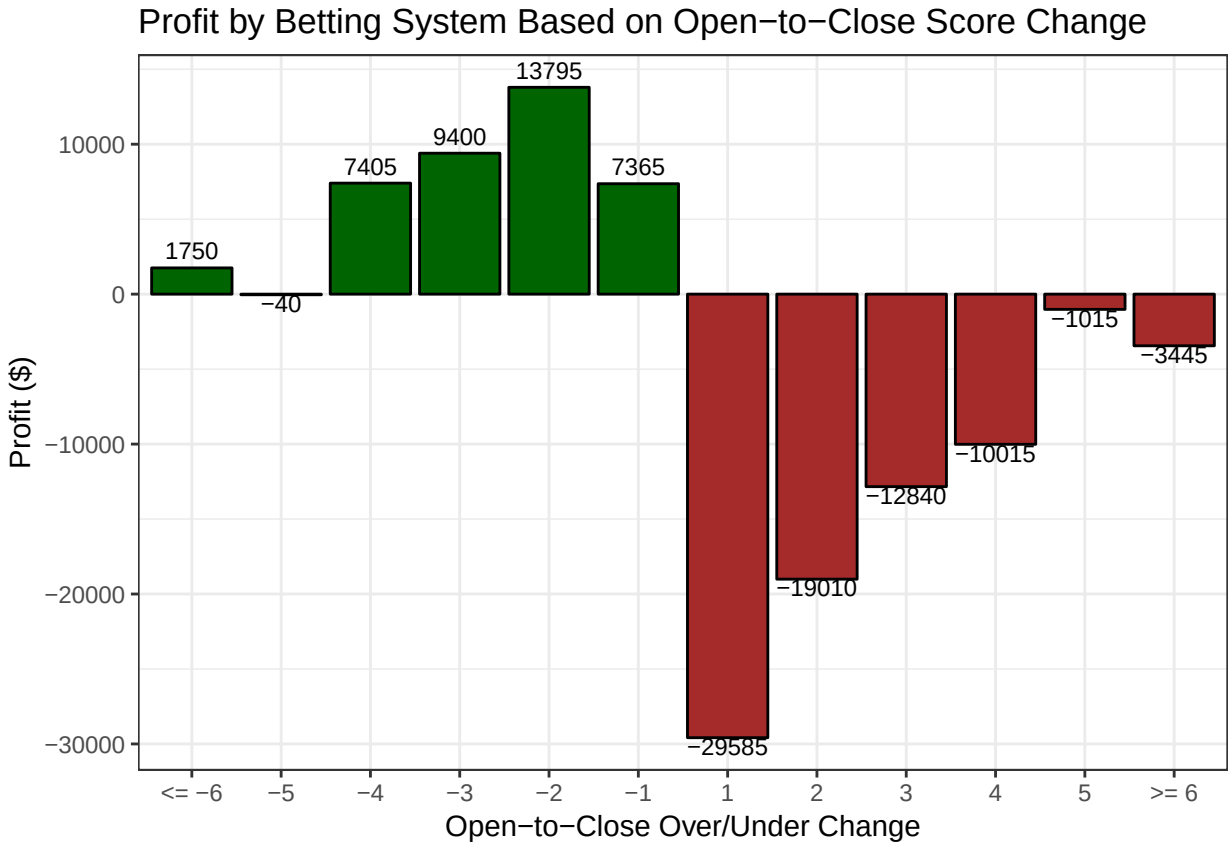
Our suspicion is that if the over/under increases (that is, more people think that the total number of points scored will be greater than first posited by the sportsbooks), this is a sign of new knowledge. As such, we will implement the betting system that first identifies which direction the over/under moved, and then either:

- bets on the second half "OVER" if the over/under line increased (OUClose-OUOpen > 0), or;
- bets on the second half "UNDER" if the over/under line decreased (OUClose-OUOpen < 0), or;

- bet nothing if the over/under line did not move between the open and the close.

Our simulation will places $100 bets in each of those instances. The following bar chart shows the net profits for each of the different line movements by implementing this betting system over the entirity of our dataset.

```
df1 <- tidy %>%
  select(OUOpen, OUClose, Total, OU2H, Total_2H) %>%
  mutate(Diff = OUClose-OUOpen) %>%
  mutate(Diff_dec_idx =
           cut(Diff,
               breaks = c(-Inf, -5.49, -4.49, -3.49, -2.49, -1.49, -0.49,
                          0.49, 1.49, 2.49, 3.49, 4.49, 5.49, Inf),
               labels = c("<= -6","-5","-4", "-3",
                          "-2", "-1", "skip",
                          "1", "2", "3",
                          "4", "5", ">= 6"))) %>%
  filter(!(Diff_dec_idx == "skip")) %>%
  mutate(earning = if_else( (Total_2H-OU2H)*(Diff)>0, 95,
                            if_else(Total_2H==OU2H, 0, -100) ) )

df2 <- df1 %>%  group_by(Diff_dec_idx) %>%
  summarise(profit = sum(earning)) %>%
  mutate( Gain = if_else(profit>0, "+", "-"))

ggplot(df2, aes(x=Diff_dec_idx, y = profit, fill = Gain)) +
  geom_bar(stat='identity', color = "black") +
  xlab("Open-to-Close Over/Under Change") +
  ylab("Profit ($)") +
  scale_fill_manual(values=c("brown", "darkgreen")) +
  geom_text(aes(label = paste(profit), vjust=if_else(profit>0,-0.5,1)), size = 3) +
  ggtitle("Profit by Betting System Based on Open-to-Close Score Change") +
  theme_bw() +
  theme(legend.position = "none")
```



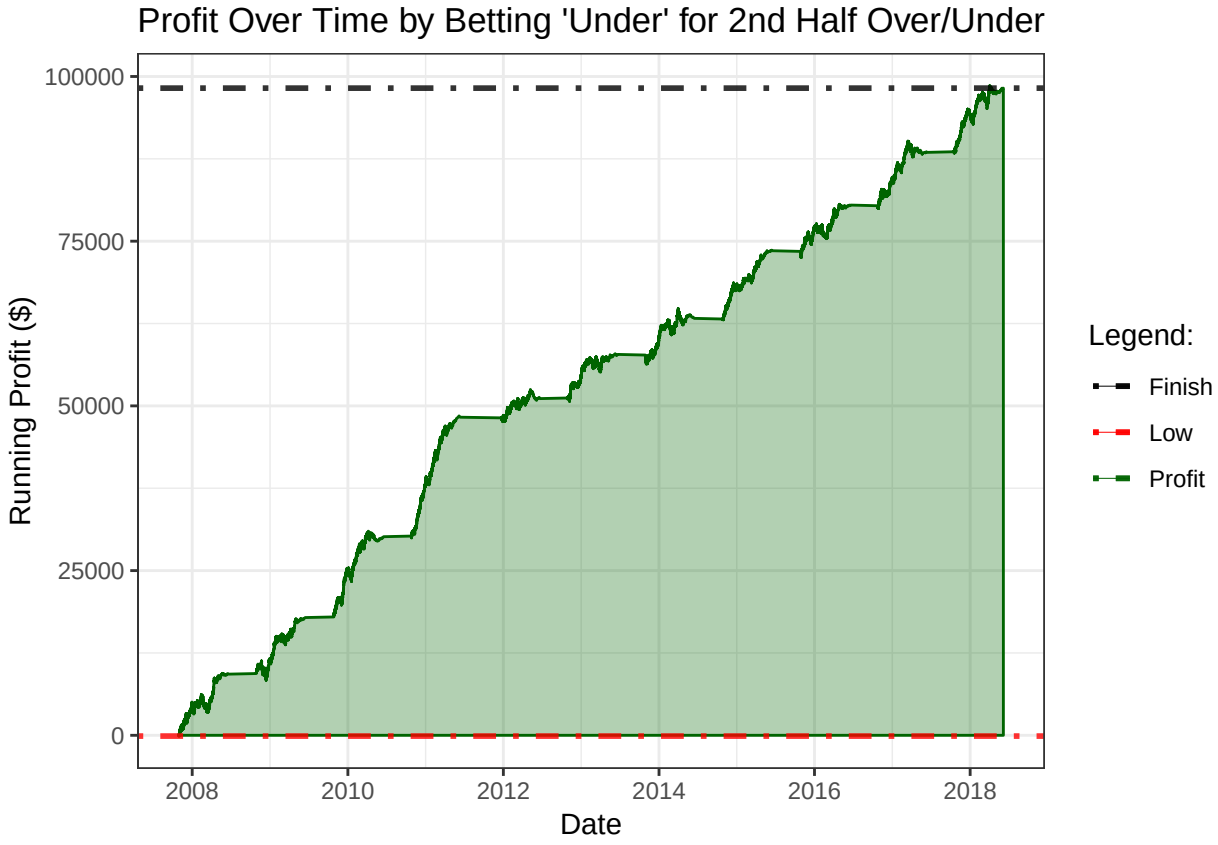
Clearly, there are some noticable trends here. When betting on the "under" for the second half when the closing over/under value was smaller than the opening over/under value, we are nearly always profitable. Conversely, when betting on the "over" for the second half in the opposite scenario, we were always unprofitable.

4.1.3 Modified Model

Logically, we will simply modify our betting system such that we would ALWAYS bet on the under for the second half, regardless of which direction the open-to-close over/under line moved. This would invert the bars on the above graph for instances where the open-to-close over/under difference was positive. Our original hypothesis has been deemed incorrect, but we have stumbled upon some interesting results. We hereby present the running profit over time by implementing a betting system that always bets the under for the second half score:

```
df <- tidy  %>% select(Date, OUOpen, OUClose, OU2H, Total_2H) %>%
  mutate(Diff = OUClose-OUOpen)

df$Profit = 0
for (i in 2:length(df$Profit)){

  if (df$OU2H[i] > df$Total_2H[i]) {
    df$Profit[i] <- df$Profit[i - 1] + 95
  }
  else if (df$OU2H[i] < df$Total_2H[i]) {
    df$Profit[i] <- df$Profit[i - 1] - 100
  }
  else {
    df$Profit[i] <- df$Profit[i - 1]
```

```
  }
}

winner <- ggplot(df, aes(y = Profit, x = Date, col = "Profit")) +
  geom_line(lwd = 0.1) +
  geom_ribbon(aes(x = Date, ymax = Profit), ymin = 0, alpha=0.3,
              fill = "darkgreen", color = "darkgreen") +
  ggtitle("Profit Over Time by Betting 'Under' for 2nd Half Over/Under") +
  labs(x = "Date", y = "Running Profit ($)") +
  theme_bw()

winner +
  geom_hline(aes(yintercept = min(Profit), colour = "Low"),
                    alpha = 0.8, lwd = 1, linetype="dotdash") +
  geom_hline(aes(yintercept = Profit[14186], colour = "Finish"),
             alpha = 0.8, lwd = 1, linetype="dotdash") +
  scale_color_manual(values = c('Profit' = 'darkgreen', "Low" = "red",
                                "High" = "blue", "Finish" = "black")) +
  labs(x = "Date", y = "Running Profit ($)", color = "Legend:")
```



Frankly, this finding is absolutely remarkable. We have re-run and re-tested this result for several hours to confirm its validity. In short, this graphic shows that if you were to have been $100 on the "under" for the second half over/under in all NBA games since the beginning of the 2007-2008 NBA season, you would be up over $98,000.

4.2 Spread Analysis

Spread betting is similar to over/under betting, in the sense that they are both theoretically 50%-50% bets. The "spread" is defined as the minimum number of points one team has to win by, or maximum number of points one team can lose by, for a bettor to win their bet. A team with a negative spread is the favorite, while the team with the positive spread is the underdog. To give a concrete example, a spread of *-5.5* indicates that bet on the favorite spread will only win if the team wins by 6 or more points. Conversely, bets on the underdog will pay out as long as the underdog does not lose by more than 5.5 points.

Another parallel to over/under betting is that the actual value of the spread is both a factor of the bookmaker's models and the market. Bookmakers publish initial spread values that can be modified based on where the market is placing their bets. Spread betting is also subject to the standard commissions, like over/under betting, which are reflected in our visualizations.

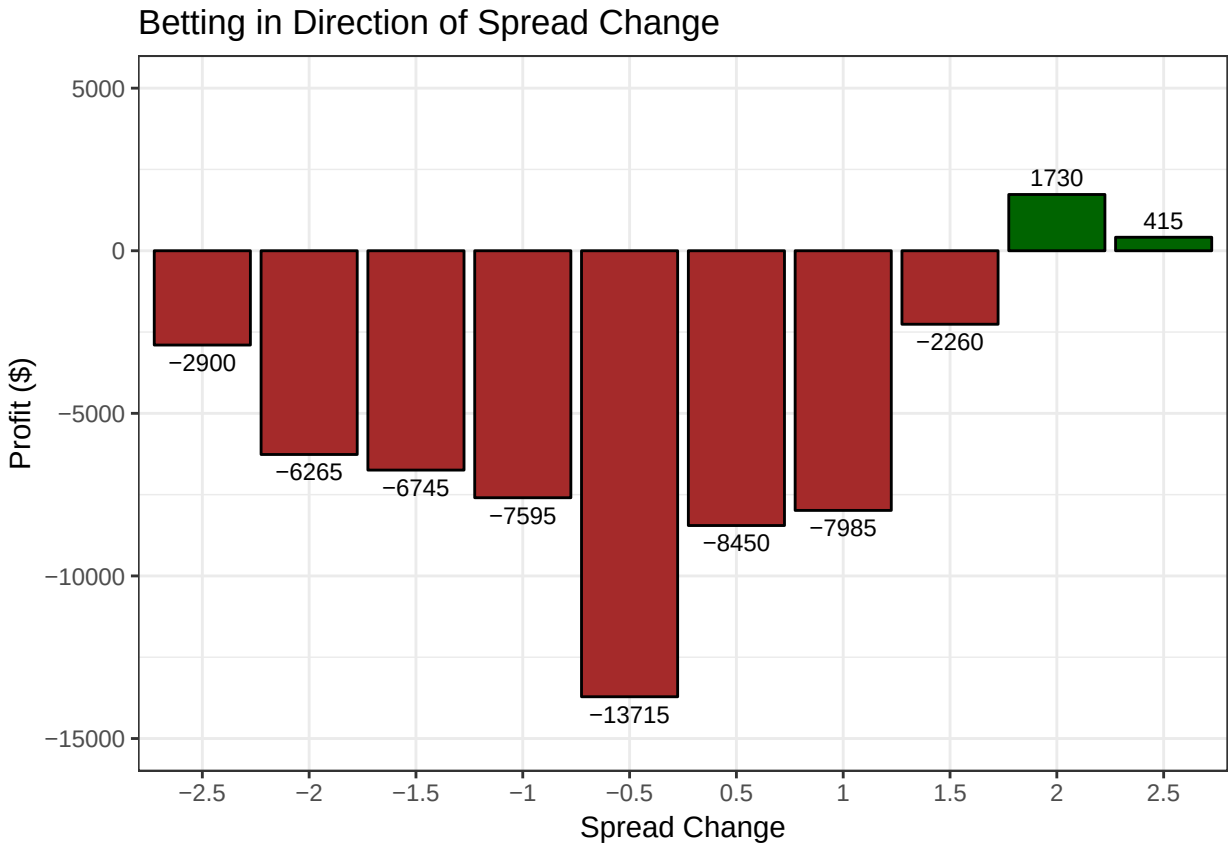
4.2.1 Spread Movement as Indicator

To see if we could replicate the success of our over/under betting model, we wanted to examine if changes in spread values between the open and close would be indicators of new information. Rather than betting on second half spreads, though, we would simply select the closing spread based on the direction that the spread line moved.

To be specific, we would calculate the spread change, which occurred in increments of 0.5 points. For this visualization, we show the net earnings over time if you always bet on a team with a given net spread change. So for example, in the visualization below, we can see that the bar corresponding to a spread change of +2 had a net profit of $1730. This means that for all the instances where a team became LESS favored by 2 points (e.g. spread changed from +2 to +4, or -7 to -5), if you bet on that team, you would have seen those net earnings over the past 11 seasons.

```
tidy <- read_csv("../Data/tidy.csv")
tidy$HSpreadChange <- tidy$HSpreadClose - tidy$HSpreadOpen
tidy$VSpreadChange <- tidy$VSpreadClose - tidy$VSpreadOpen
tidy <- tidy %>% select(Date, VF, HF, HSpreadClose,
                        VSpreadClose, HSpreadChange, VSpreadChange)

spreadChange <- data_frame("Change" = c(-2.5, -2, -1.5, -1,
                                        -0.5, 0.5, 1, 1.5, 2, 2.5))
spreadChange$Profit <- 0

for (j in 1:length(spreadChange$Change)){
  tidy$Payout <- 0
  for (i in 1:length(tidy$Payout)) {
    if (spreadChange$Change[j] == tidy$HSpreadChange[i]) {
      if (tidy$VF[i] - tidy$HF[i] < tidy$HSpreadClose[i]) {
        tidy$Payout[i] <- 95
      } else if (tidy$VF[i] - tidy$HF[i] > tidy$HSpreadClose[i]){
        tidy$Payout[i] <- -100
      }
    } else if (spreadChange$Change[j] == tidy$VSpreadChange[i]) {
      if (tidy$HF[i] - tidy$VF[i] < tidy$VSpreadClose[i]) {
        tidy$Payout[i] <- 95
      } else if (tidy$HF[i] - tidy$VF[i] > tidy$VSpreadClose[i]) {
        tidy$Payout[i] <- -100
      }
    }
  }
```

```
  spreadChange$Profit[j] = sum(tidy$Payout)
}
spreadChange <- spreadChange %>%  mutate( Gain = if_else(Profit>0, "+", "-"))
spreadChange$Change <- as.factor(spreadChange$Change)
com <- ggplot(spreadChange, aes(x=Change, y=Profit, fill  = Gain)) +
  geom_bar(stat='identity', color='black') +
  geom_text(aes(label = paste(Profit),
                vjust = ifelse(Profit >= 0, -0.5, 1.5)), size=3) +
  scale_y_continuous(limits = c(-15000,5000)) +
  scale_fill_manual(values=c("brown", "darkgreen")) +
  labs(x="Spread Change", y = "Profit ($)",
       title="Betting in Direction of Spread Change") +
  theme_bw() +
  theme(legend.position = "none")

com
```



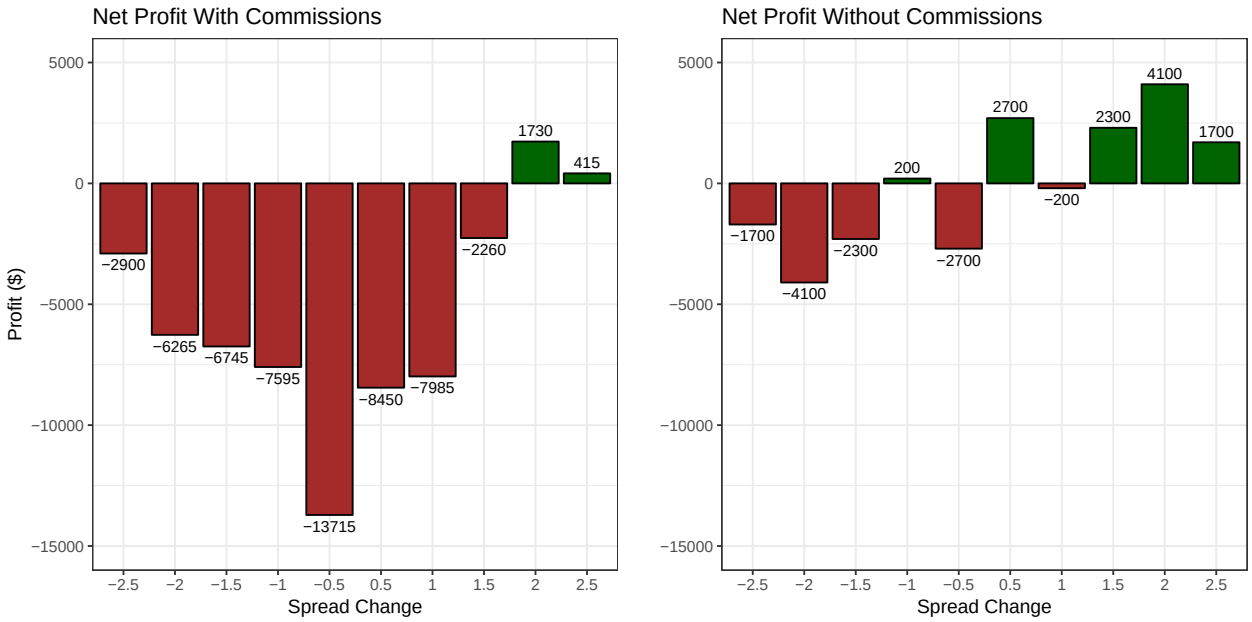
This, on the surface, does not appear to be a valuable source of information for bettors. Regardless of whether you made bets in favor or against the direction that the spread line moved, you would have walked away a loser at the end of the day for nearly all spread change values.

4.2.2 A No Commission World

This result fact, though, gives us a great opportunity to highlight the damaging long term effect of paying a 2.5% commission on every bet, as was done in this simulation. To showcase this, we will present the outcomes of running the exact same simulation, but this time, winning $100 on a $100 bet, which in effect serves the

purpose of operating in a zero commission world.

```
for (j in 1:length(spreadChange$Change)){
  tidy$Payout <- 0
  for (i in 1:length(tidy$Payout)) {
    if (spreadChange$Change[j] == tidy$HSpreadChange[i]) {
      if (tidy$VF[i] - tidy$HF[i] < tidy$HSpreadClose[i]) {
        tidy$Payout[i] <- 100
      } else if (tidy$VF[i] - tidy$HF[i] > tidy$HSpreadClose[i]){
        tidy$Payout[i] <- -100
      }
    } else if (spreadChange$Change[j] == tidy$VSpreadChange[i]) {
      if (tidy$HF[i] - tidy$VF[i] < tidy$VSpreadClose[i]) {
        tidy$Payout[i] <- 100
      } else if (tidy$HF[i] - tidy$VF[i] > tidy$VSpreadClose[i]) {
        tidy$Payout[i] <- -100
      }
    }
  }
  spreadChange$Profit[j] = sum(tidy$Payout)
}

spreadChange <- spreadChange %>%  mutate( Gain = if_else(Profit>0, "+", "-"))
spreadChange$Change <- as.factor(spreadChange$Change)
nocom <- ggplot(spreadChange, aes(x=Change, y=Profit, fill = Gain)) +
  geom_bar(stat='identity', color='black') +
  geom_text(aes(label = paste(Profit),
                vjust = ifelse(Profit >= 0, -0.5, 1.5)), size=3) +
  scale_y_continuous(limits = c(-15000,5000)) +
  scale_fill_manual(values=c("brown", "darkgreen")) +
  labs(x="Spread Change", y = " ",
       title="Net Profit Without Commissions") +
  theme_bw() +
  theme(legend.position = "none")

com <- com + labs(title = "Net Profit With Commissions")
grid.arrange(com, nocom, ncol = 2)
```



The difference is clear. For all of the spread change values, net profits are significantly greater (or losses significantly less). The symmetry for positive and negative spread change values also reveals itself. This makes logical sense, because if you bet in favor of the spread line movement (negative spread change bars) for one team, you would bet for their opponent to beat the spread when you bet against the spread line movement (positive spread change bars).

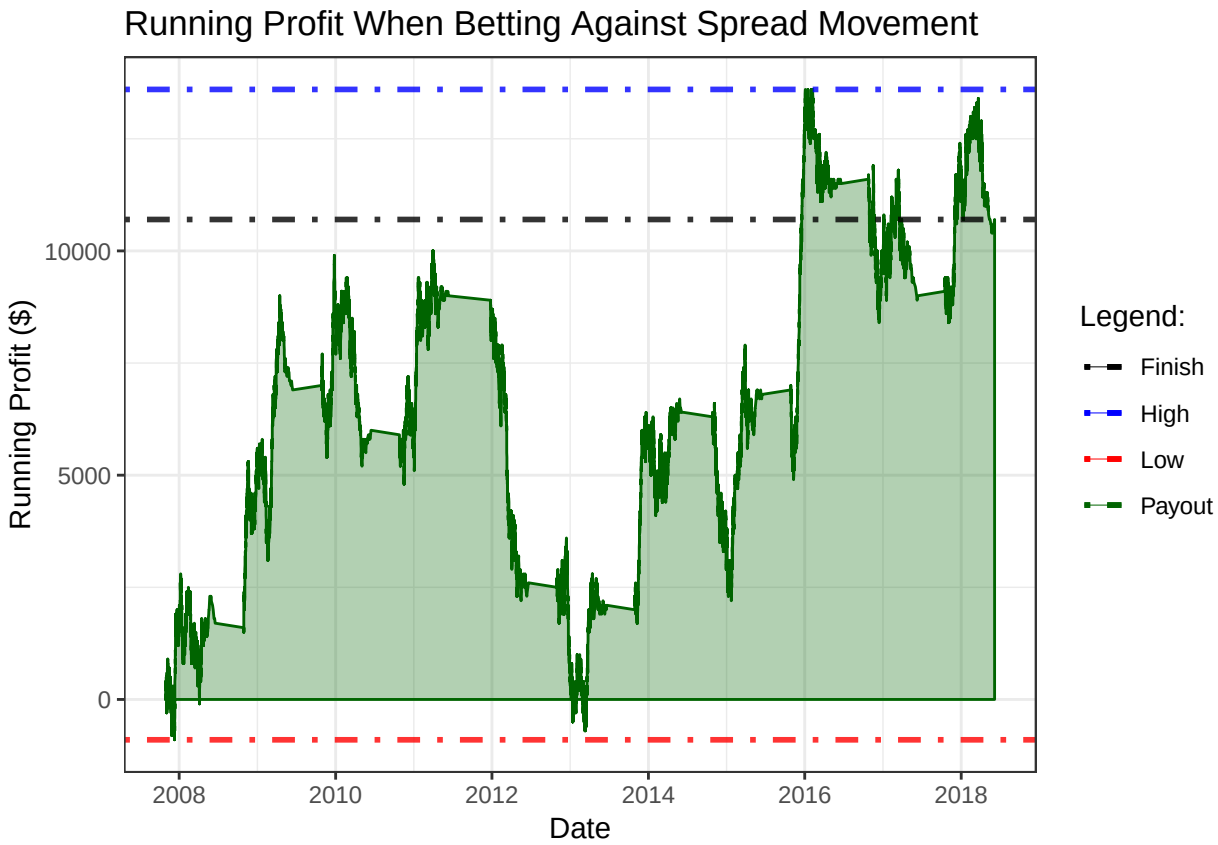
Now we can see that it is clearly in one's interest to bet against the movement of the spread if it is possible to remove the commission. To give a concrete example. If a team of interest was initially an underdog needing to beat a +5 spread at the open, and then the line moved to a +7 spread at the close (close - open = 7 - 5 = +2), then you should bet FOR that team to beat the +7 spread.

4.2.3 No Commission, Running Profit

To highlight this disparity, we wanted to show a running profit over time by using the betting system described above in a zero commission world. This is not entirely unrealistic, as bettors that are set up with multiple sportsbooks will have a selection of odds to choose from. So, while the lines between sportsbooks are very similar, there are often disparities that can be taken advantage of on the part of the bettor, in order to effectively reduce the commission that they pay on every bet.

```
tidy$Payout <- 0
tidy$Payout[1] <- 100
for (i in 2:length(tidy$Payout)) {
  if (tidy$HSpreadChange[i] > 0) {
    if (tidy$VF[i] - tidy$HF[i] < tidy$HSpreadClose[i]) {
      tidy$Payout[i] <- tidy$Payout[i - 1] + 100
    } else if (tidy$VF[i] - tidy$HF[i] > tidy$HSpreadClose[i]){
      tidy$Payout[i] <- tidy$Payout[i - 1] - 100
    } else {
      tidy$Payout[i] <- tidy$Payout[i - 1]
    }
  } else if (tidy$VSpreadChange[i] > 0) {
    if (tidy$HF[i] - tidy$VF[i] < tidy$VSpreadClose[i]) {
      tidy$Payout[i] <- tidy$Payout[i - 1] + 100
    } else if (tidy$HF[i] - tidy$VF[i] > tidy$VSpreadClose[i]) {
```

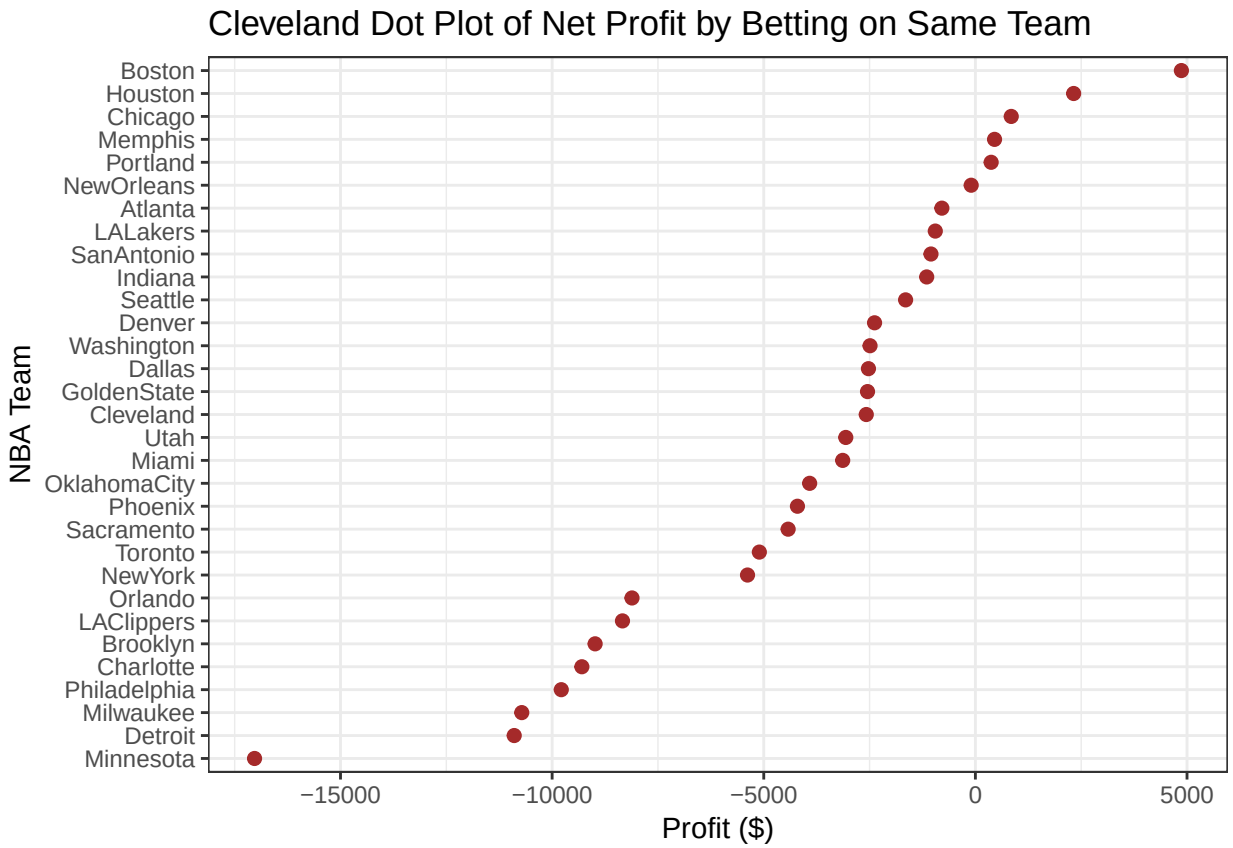
```
      tidy$Payout[i] <- tidy$Payout[i - 1] - 100
    } else {
      tidy$Payout[i] <- tidy$Payout[i - 1]
    }
  } else {
      tidy$Payout[i] <- tidy$Payout[i - 1]
  }
}

ggplot(tidy, aes(y = Payout, x = Date, col = "Payout")) +
  geom_line(lwd = 0.1) +
  geom_ribbon(aes(x = Date, ymax = Payout), ymin = 0, alpha=0.3,
              fill = "darkgreen", color = "darkgreen") +
  geom_hline(aes(yintercept = max(Payout), colour = "High"),
             alpha = 0.8, lwd = 1, linetype="dotdash") +
  geom_hline(aes(yintercept = min(Payout), colour = "Low"), alpha = 0.8, lwd = 1,
             linetype="dotdash") +
  geom_hline(aes(yintercept = Payout[14186], colour = "Finish"), alpha = 0.8,
             lwd = 1, linetype="dotdash") +
  ggtitle("Running Profit When Betting Against Spread Movement") +
  scale_color_manual(values = c('Payout' = 'darkgreen', "Low" = "red",
                                "High" = "blue", "Finish" = "black")) +
  labs(x = "Date", y = "Running Profit ($)", color = "Legend:") +
  theme_bw()
```



4.3 Individual Teams

As a last visualization, we thought that it would be interesting to see how fans of specific teams have performed over time. We hypothesized that some factors could be relevant in regards to which teams have their moneylines (effectively, probability of winning a game) overvalued or undervalued. For example, since New York has a much larger population than Memphis, this may result in more people betting on the New York Knicks to win the game in a straight moneyline bet than the Grizzlies. If this is the case, New York would be disproportionately overvalued, resulting in lower payouts than the corresponding chance of them winning a game.

```
tidy <- read_csv("../Data/tidy.csv")
df2 <- tidy %>%
  select(V, H, VF, HF, VMoney, HMoney) %>%
  mutate(Bet_V_win = if_else(VF>HF, if_else(VMoney>0, VMoney, -100/VMoney*100),
                             if_else(VF<HF, -100, 0))) %>%
  mutate(Bet_H_win = if_else(VF<HF, if_else(HMoney>0, HMoney, -100/HMoney*100),
                             if_else(VF>HF, -100, 0)))

df_V <- df2 %>%
  group_by(V) %>% summarise_at("Bet_V_win", sum) %>%
  rename(Team = V)

df_H <- df2 %>%
  group_by(H) %>% summarise_at("Bet_H_win", sum) %>%
  rename(Team = H)

df_Team <- df_V %>%
  inner_join(df_H) %>%
  mutate(profit = Bet_V_win + Bet_H_win)

ggplot(df_Team, aes(x= profit, y= fct_reorder(Team, profit))) +
  geom_point(color = "brown", size = 2) +
  labs(x="Profit ($)", y="NBA Team") +
  ggtitle("Cleveland Dot Plot of Net Profit by Betting on Same Team") +
  theme_bw()
```



From this Cleveland Dot Plot, it is evident that some teams have had more success than others on the betting markets. The exact patterns or reasoning is not evident, though. From this graph we can also see the commissions playing a factor, as only five teams would have resulted in a net profit that is positive, if a bettor would have exclusively bet on this team. In a zero commission world, these dots would be centered approximately around a betting profit of $0.

This concludes our main analysis. If you would like an opportunity to discover your own betting system, please see our interactive component! From there, you can select your favorite NBA team and a particular betting method, and immediately discover how you would have performed in the betting markets over the past 11 seasons!

Link: http://www.fernandotroeman.com/nbaodds/

5 Executive Analysis

5.1 Introduction

5.1.1 Theory Background

When one walks into a casino, one can precisely state the probability of winning and losing at certain games. This is because all variables at play have been calculated beforehand, and the payouts for winning/losing are determined accordingly such that by the law of large numbers, the gambler will always lose to the house.

Sports matches, on the other hand, consists of an infinite number of variables which can only be estimated, and never be known for certain. But, since sportsbooks will offer bets on the outcomes of matches, this necessitates that they must gauge the probability of the outcomes in order to offer a fair bet. However, since these probabilities are only estimates, there may be a discrepancy between the probability of winning certain bets and the TRUE probability (incalculable in theory) of those events occuring. If such differences are found, this leaves an opening for potential exploitation.

5.1.2 Sports Betting Background

Sports betting offers many different avenues by which bets can be made. The most popular such bets are for the outright game winner, the total amount of points in the game, and the margin of victory between two teams.

In this report, we specifically focused on betting pertaining to NBA basketball games. We investigated possible betting systems for all of the aforementioned betting avenues, and attempted to discover systems that would have been profitable if implemented over the course of the past 11 NBA seasons. If such a system exists, it would reveal flaws within the betting market's ability to accurately determine the probabilities of certain events, which could in turn be exploited for capital gain.

5.2 Findings

5.2.1 Betting System Description

Our most successful betting system pertains to over/under betting for the second half of basketball games. In such a bet, a sportsbook will provide a **number.** The bettor can then bet on whether they think the total number of points scored in the second half of the basketball game will be *over* or *under* this **number.** The **number** is chosen by the sportsbook such that they estimate that there will be a 50%-50% chance of the true number of points scored in the game being above or below this **number**. We wanted to see whether these numbers were indeed chosen such that 50% of games totaled more points and 50% totaled less points.

Through a gradually evolving betting system that built upon previous discoveries, we were able to finalize our very simple, yet extremely effective, betting system. In the first iteration of the discovery process, we proposed a betting system that would pick the over or the under based on prior information. The prior information turned out to be irrelevant, but regardless, it lead us in the direction that we needed to go. The results are revealed in Section 5.2.2.

5.2.2 Outcomes

In Figure 1, you will see the net profit for each year since the 2007 NBA season by betting $100 on every NBA games's "under" side for the "over/under" bet. In Figure 2, please find a graph showing the running profit over this time period.

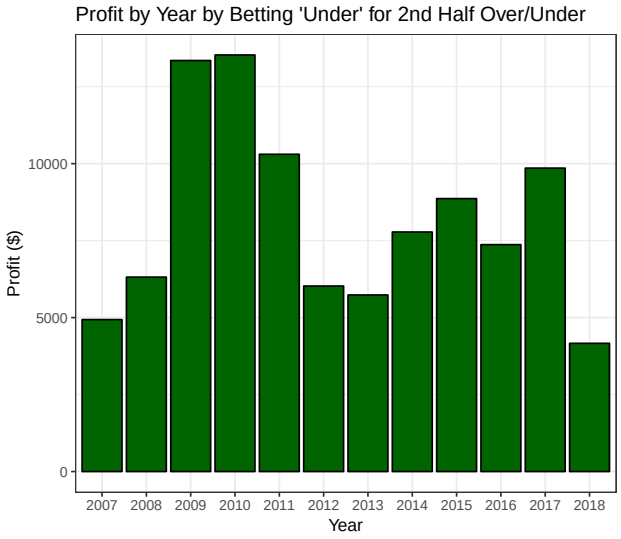


Fig. 1

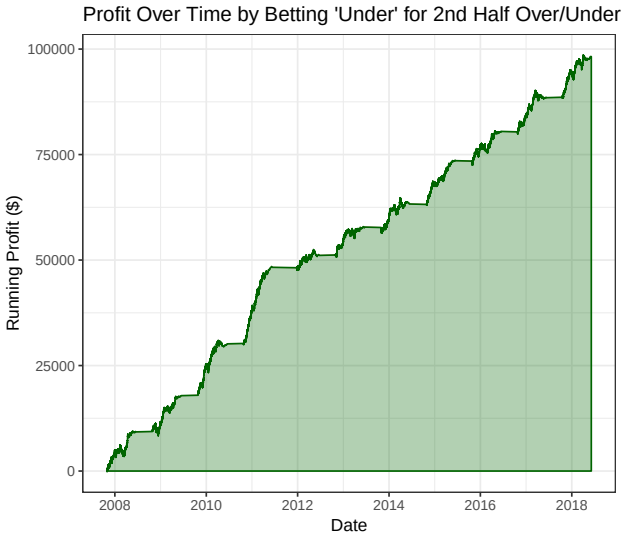


Fig. 2

5.3 Summary

These results are quite significant. They show that our proposed betting system would have made a bettor nearly $100,000 if it had been implemented since 2007. Given that 14186 games were played during this time span, this corresponds to a greater than 7% advantage for every single bet that is placed.

In essence, this is equivalent to offering a "double your money" wager for an event that occurs 57% of the time. What makes this finding even more meaningful, is the consistency of the returns year-over-year. As can be seen in Figure 1, no year is unprofitable. Although no statistical analysis have been made in regards to significance, the sheer sample size makes it clear that no such test is needed. If the consistency of this model had been known during this actual time period, a betting system that changed bet amounts based on the current bankroll would have been much more profitable, as it would not have been limited to insignificant $100 bets when the bankroll had increased by tens of thousands of dollars.

In statistical terms, the key finding here is that the sportsbooks overestimate the number of points that will be scored in the second half of NBA games. Because of this, the amount of games that have fewer points scored than the over/under value provided by the sportsbooks is much greater than the amount of games that score more points than the over/under value. Since the payouts correspond to a 50%-50% wager, though, this makes the under-side very profitable in the long run.

With that said, it is of course impossible to predict future games. However, there cannot be a stronger indicator of future results than past results in this domain!

6 Conclusion

6.1 Limitations

The greatest limitation with regards to this report as it pertains to the results, in our estimation, is the data. Error can seep into every dataset, and quality control over 14000 matches with 27 columns of variables is difficult to perform on such a scale. With that said, we did not think that this was enough of a concern to abandon our search for patterns and betting systems in this very interesting dataset. We performed a thorough and fairly comprehensive data cleaning to mitigate the effect of possible mis-entries, while frequently confirming data entries along the way. It was a very healthy exercise in data exploration and visualization from top to bottom.

6.2 Lessons Learned

From a sports betting standpoint, we learned a lot. First and foremost, we were shocked at the accuracy of the models implemented by sportsbooks to gauge probabilities of outcomes. Although we highlighted the most significant findings, which indeed were extremely intriguing, there were many more betting systems where the sportsbooks squeezed out any margin for profit with their commissions. In the long run, to be profitable as a bettor, one must find a model that has an advantage that exceeds the commissions that they pay (which tend to vary from 2-4%), and that is extremely difficult.

As discussed, we also learned that it is indeed possible to be profitable if the right system is found. It is, however, only a matter of time before such market inefficiencies are discovered and balanced out (in the case of second half over/unders, the sportsbook-given over/under value will inevitably decrease). It is impossible to know when this transition takes place, and as such, bettors are susceptible to losing by implementing a model that is no longer valid.

From a data analysis and exploration standpoint, this project was incredibly educational. The ugly sides of data science were clear with processes such as data cleaning, failed models, and uncooperative technology, et cetera. However, the process of discovering something truly remarkable was very rewarding.

6.3 Future Directions

With regards to this exact dataset, and the general practice of trying to find profitable betting systems, we have certainly discussed ideas as a group of how to proceed. There are many data sources out there, and after our tuition bills come in the mail, it would only be irresponsible of us to avoid exploring avenues by which it is possible to make a few extra dollars! A few systems we have discussed exploring are:

- The relationship between the over under and the moneyline
- Investigate how previous season champions perform against expectations
- Could see how record at a given point in time affects how people bet
- Performance of teams in back-to-back games

THE END :)