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 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 overly stimulated by:

1) Sports
2) Money
3) Gambling

. . . and of course:

4) Data exploration and visualization

We therefore found this data set to be an excellent suitor for delivering captivating insights and exercising our data analysis and visualization skills

More specifically, we are particularly interested in asking (and hopefully definitively 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 the code can be found in our GitHub document 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, which then had to be merged to a single row during the tidying process. The following columns were included in the raw, untidy dataset:

1) Date - Given in the integer format “MMDD”
2) V/H - 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, respectively
5) Final - Final score for the full game *Note - 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 an optional, but comprehensive, section here that 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 theoretically a 50%-50% bet (being correct/incorrect should have the same probability as for a coin flip). 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, the selected team must win by MORE than the spread value if they are the favorite, or the selected team 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 sacrificed 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.

3 Analysis of Quality

3.1 Tidying

The data, once merged for the past 11 seasons, had to be converted into a tidy format for it to be possible to conduct any data analysis. More specifically, each individual NBA match in the dataset had separate rows for the two opposing teams. To tidy, then, we had to merge this by adding a few columns so that each data entry (row) was only comprised of one game.

Before tidying, the CSV looked like this:

```r
options(dplyr.width = Inf)
kable(combined[1:5,], 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
2007-10-30	2007	V	Houston	16	27	27	25	95	2.5	5	-230	3

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

```r
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 assessed 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 they could be easily removed later. 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'))[1:5,], 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
2007-11-03	2007	V	GoldenState	28	18	35	29	110	212.5	215.5	425	pk
2007-11-06	2007	H	NewYork	32	28	22	37	119	213	211	140	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

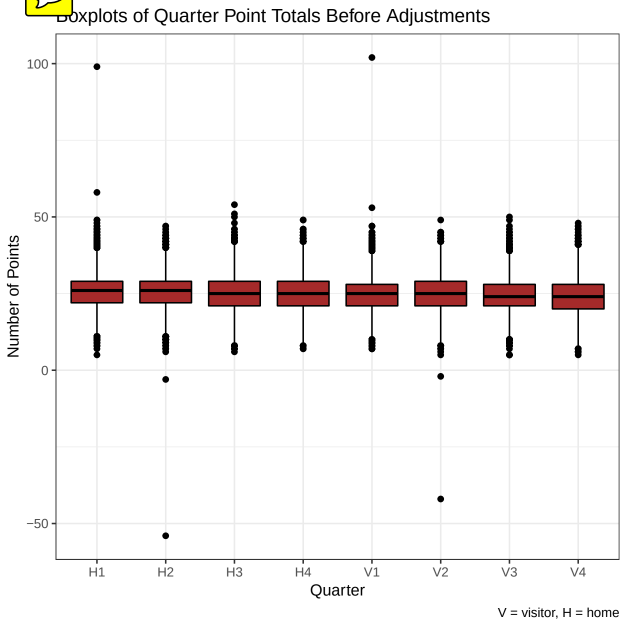
```
# Replace all pk odds (i.e. 50/50 outcomes) with 0; remove NL bets
df = df.replace(to_replace=["pk", "PK"], value=0)
df = df.replace(to_replace="NL", value=np.nan)

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

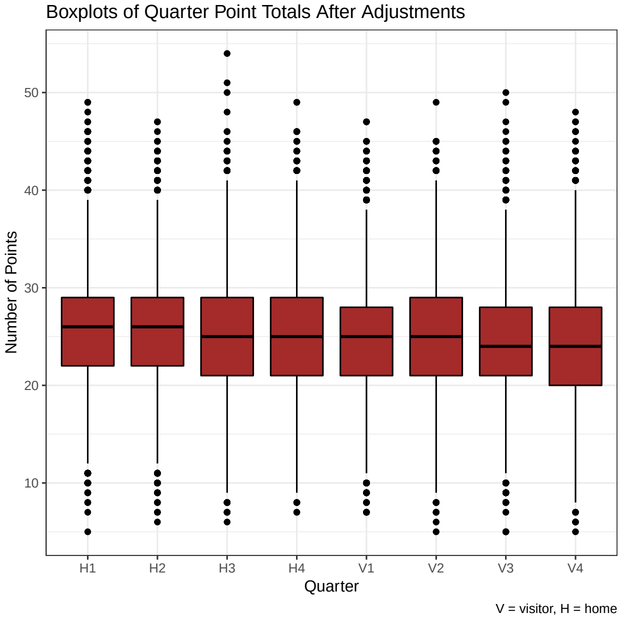
3.2.3 Bad Entries:

To located the bad entries, we found it effective to make graphs of the different variables - particularly box plots given the number of data entries. This was done 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))
```





Clearly, negative values are 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 too harmful to be rather loose with removing data entries that weren't cooperating.

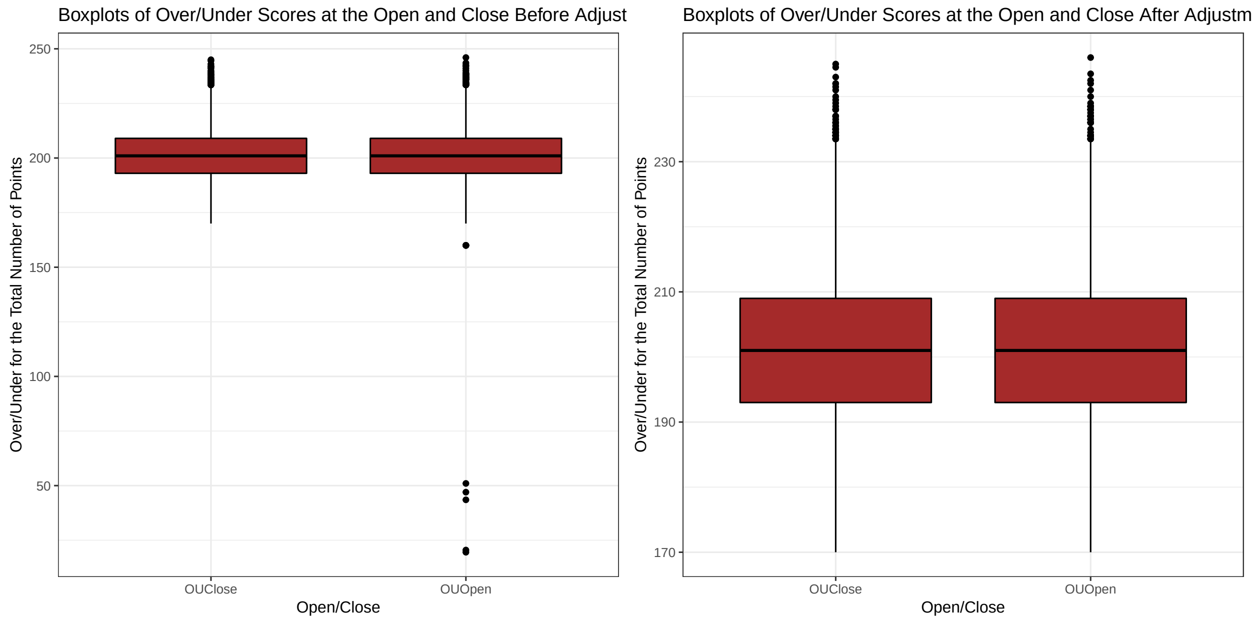
Excerpt from the clearning script in Python:

```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]
```

As can be seen, the quarter distributions looked much better after applying the corrections.

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

```r
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))
```

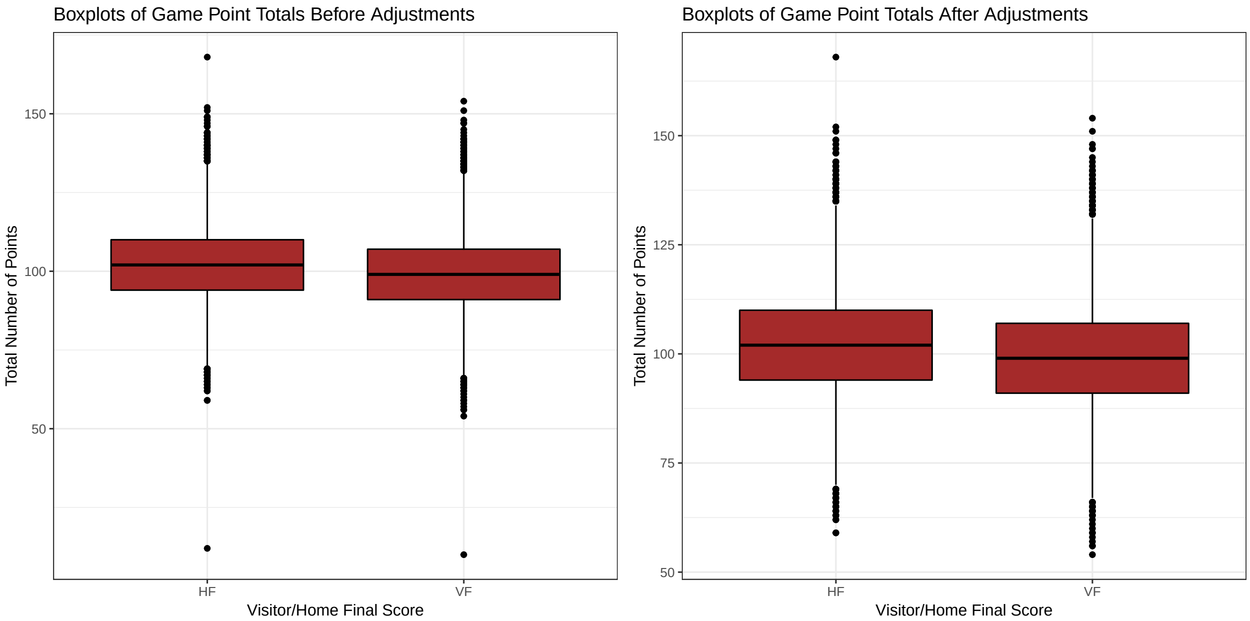


The few outliers are very obvious, so they were easily cleaned by:

```
# 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 to 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 arbitrary value. In reality, this “arbitrary value” is selected based on various predictive models by the sportsbooks, and further, is capable of changing 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).

This bet is attractive to bettors because it is, in theory, a 50%-50% bet. That is, payouts are equal regardless of which side you pick, which is not necessarily true for picking game winners with a moneyline bet. In reality, sportsbooks charge a commission, which has been reflected in the graphs that we present here, so the payouts do not exactly correspond to the probability of winning the bet (i.e. winning does not completely ‘double your money’ in this case).

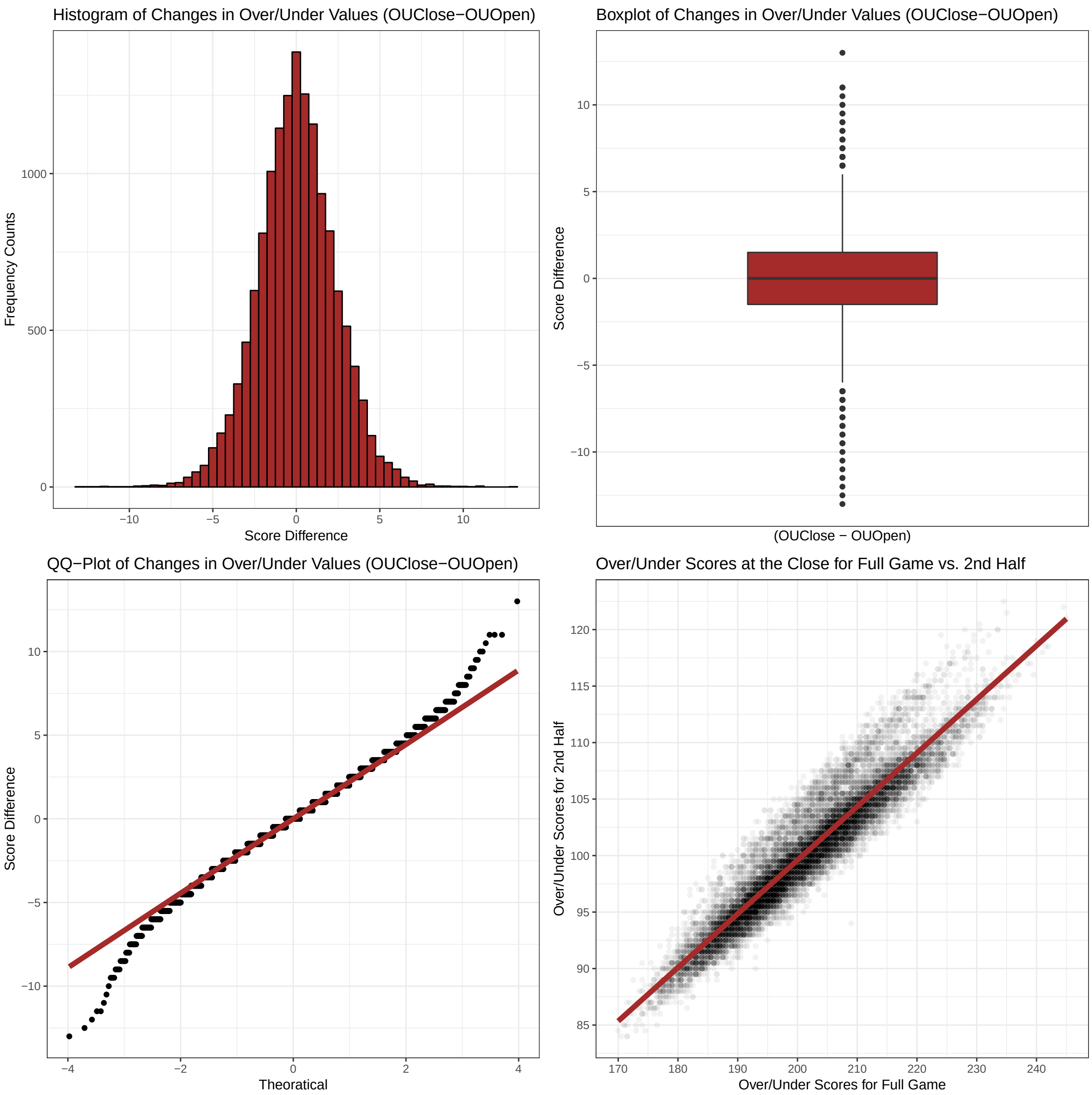
Given our intuition that changes in the over/under values carry some information (perhaps an indicator of new information regarding the game), we wanted to explore whether such swings had any relationship to the true outcomes. Additionally, we wanted to see the relatinship between the full game over/under 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.

```r
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)
```

```r
histo <- ggplot(df, aes(x = Diff)) +
  geom_histogram(binwidth = 0.5, fill='brown', color='black') +
  xlab("Score Difference") +
  ylab("Frequency Counts") +
  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 distributions 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

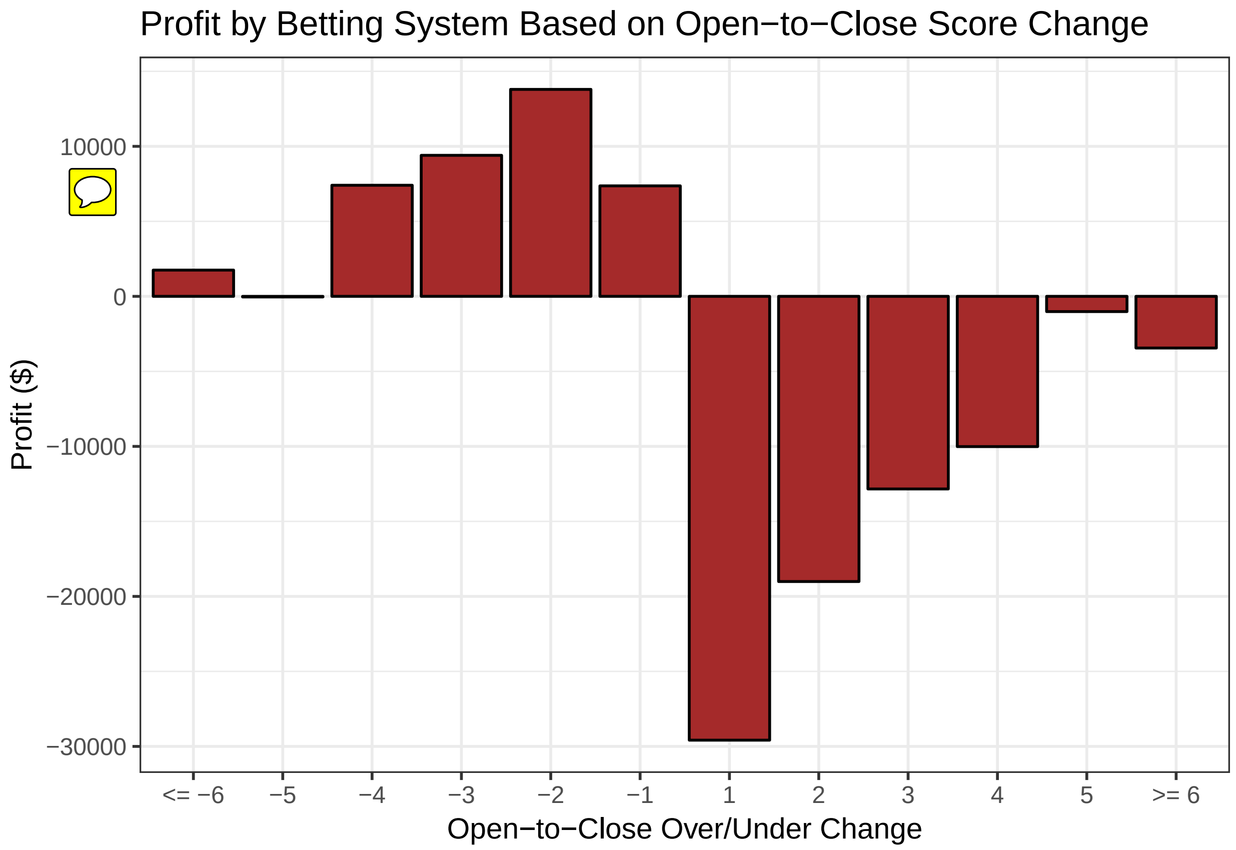
Our suspicion is that if the over/under increases (that is, more people think that the total number of points scored will be greater than first posited by the sportsbooks), this is a sign of new knowledge. As such, we will implement the betting system that first identifies which direction the over/under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

```
df <- 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 <- df %>%  group_by(Diff_dec_idx) %>%
  summarise(profit = sum(earning)) %>%
  filter(as.numeric(Diff_dec_idx)>0)

ggplot(df2, aes(x=Diff_dec_idx, y = profit)) +
  geom_bar(stat='identity', fill = "brown", color = "black") +
  xlab("Open-to-Close Over/Under Change") +
  ylab("Profit ($)") +
  ggtitle("Profit by Betting System Based on Open-to-Close Score Change") +
  theme_bw()
```

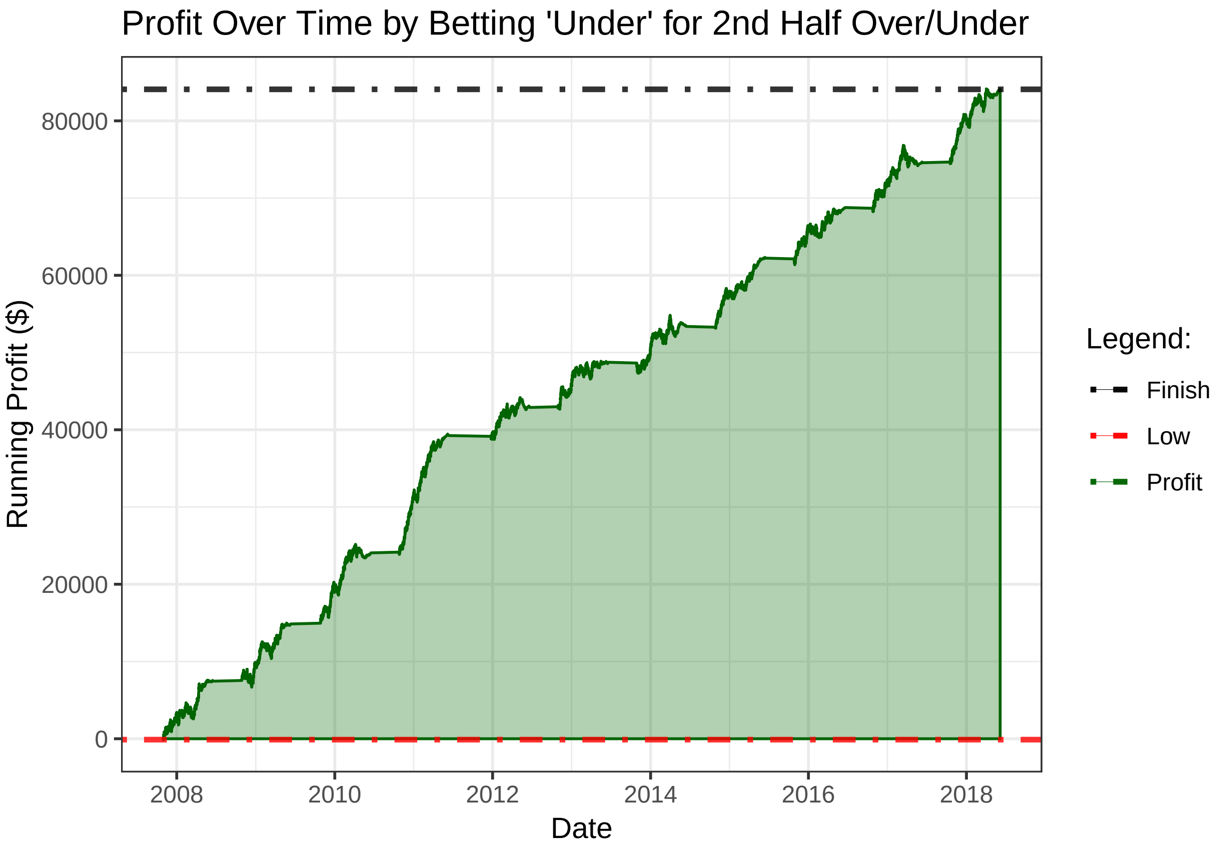


Clearly, there are some noticable trends here. When betting on the "under" for the second half when the closing over/under value was smaller than the opening over/under value, we are nearly always profitable. Conversely, when the closing over/under value was greater than the opening over/under value,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

4.1.3 Modified Model

Logically, we simply modified 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 flip the bars on the above graph for instances where the open-to-close over/under difference was positive. We have presented the running profit over time by implementing such a betting system:

```r
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$Diff[i] == 0) {
    df$Profit[i] <- df$Profit[i - 1]
  }
  else {
    if (df$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]
    }
  }
}

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") +
  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") +
  ggtitle("Profit Over Time by Betting 'Under' for 2nd Half Over/Under") +
  scale_color_manual(values = c('Profit' = 'darkgreen', "Low" = "red",
                                "High" = "blue", "Finish" = "black")) +
  labs(x = "Date", y = "Running Profit ($)", color = "Legend:") +
  theme_bw()
```



Frankly, this finding is absolutely remarkable. We have spent several hours trying to identify mistakes in our code, but we were not able to find any. In short, this graphic shows that if one 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, they would be up over $80,000.

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.

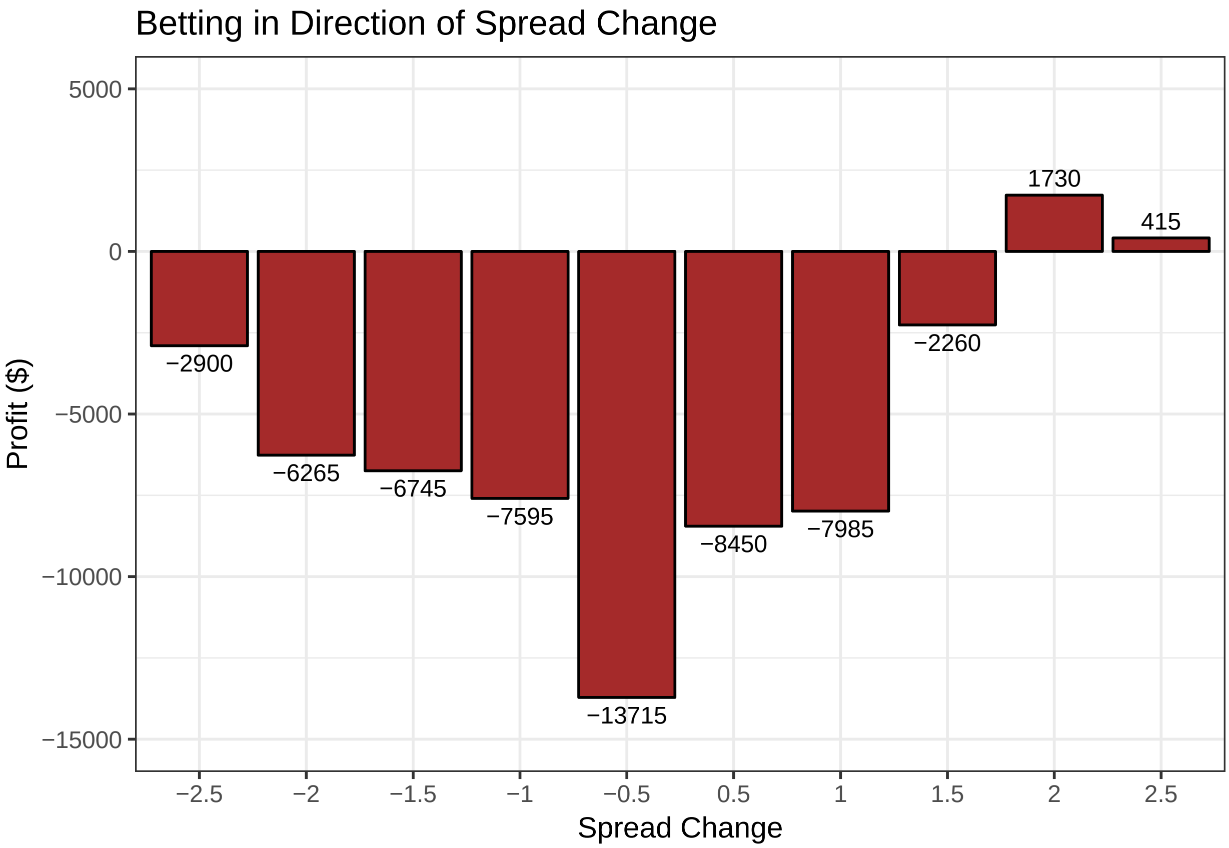
Another parallel to over/under betting is that the actual value of the spread is both a factor of the bookmaker's models and the market. Bookmakers publish initial spread values that can be modified based on where the market is placing their bets. Spread betting is also subject to the standard commissions,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

```r
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$HSpreadChange[i]) {  # if spread change for the game is a hit
      if (tidy$VF[i] - tidy$HF[i] < tidy$HSpreadClose[i]) {    # if home team beat spread
        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]) {    # if vis team beat spread
        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$Change <- as.factor(spreadChange$Change)
com <- ggplot(spreadChange, aes(x=Change, y=Profit)) +
  geom_bar(stat='identity', fill='brown', 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)) +
  labs(x="Spread Change", y = "Profit ($)", title="Betting in Direction of Spread Change") +
  theme_bw()
com
```



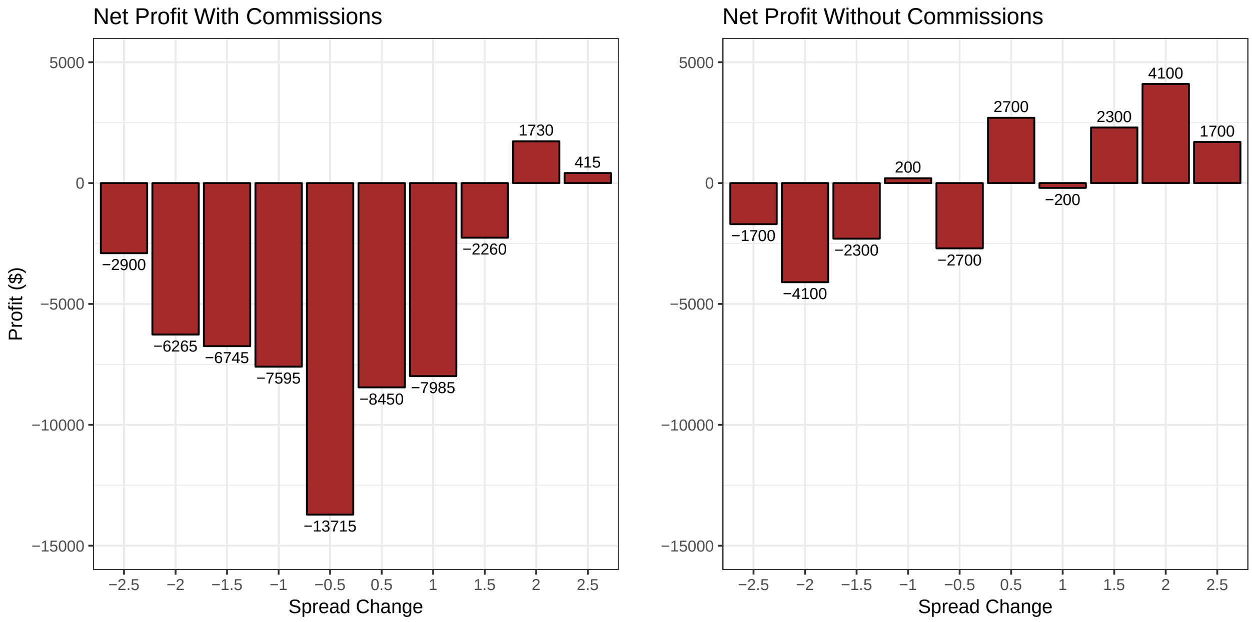
Clearly, this is not an valuable source of information for bettors. Regardless of whether you made your bets in favor or against the direction that the spread line moved, you would have walked away a loser at the end of the day for nearly all spread change values.

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$HSpreadChange[i]) {  # if spread change for the game is a hit
      if (tidy$VF[i] - tidy$HF[i] < tidy$HSpreadClose[i]) {    # if home team beat spread
        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]) {    # if vis team beat spread
        tidy$Payout[i] <- 100
      } else if (tidy$HF[i] - tidy$VF[i] > tidy$VSpreadClose[i]) {
        tidy$Payout[i] <- -100
      }
    }
```

```r
  }
  spreadChange$Profit[j] = sum(tidy$Payout)
}

spreadChange$Change <- as.factor(spreadChange$Change)
nocom <- ggplot(spreadChange, aes(x=Change, y=Profit)) +
  geom_bar(stat='identity', fill='brown', 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)) +
  labs(x="Spread Change", y = " ",title="Net Profit Without Commissions") +
  theme_bw()

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



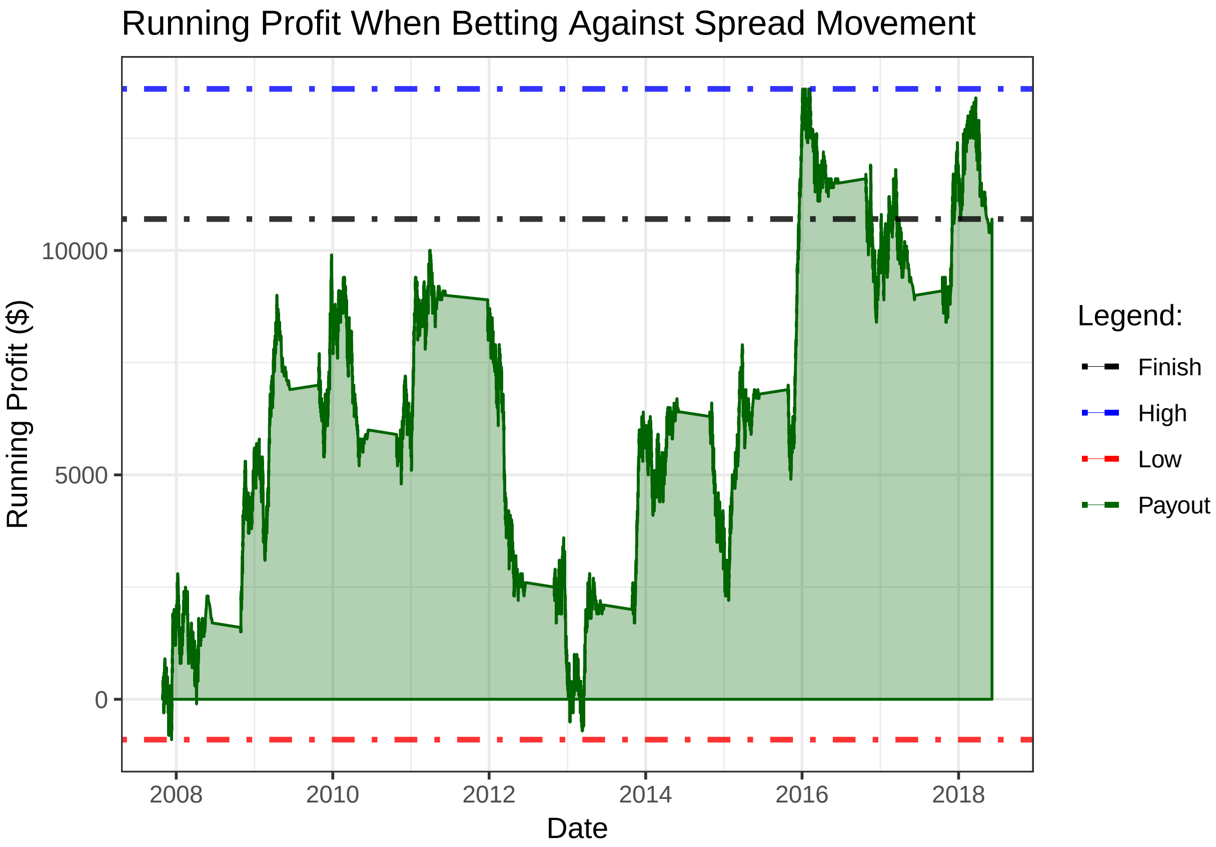
The difference is clear. For all of the spread change values, net profits are significantly greater. The symmetry for positive and negative spread change values also reveals itself. This makes logical sense, because if you would bet in favor of the spread line movement (negative spread change bars) for 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 always 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.

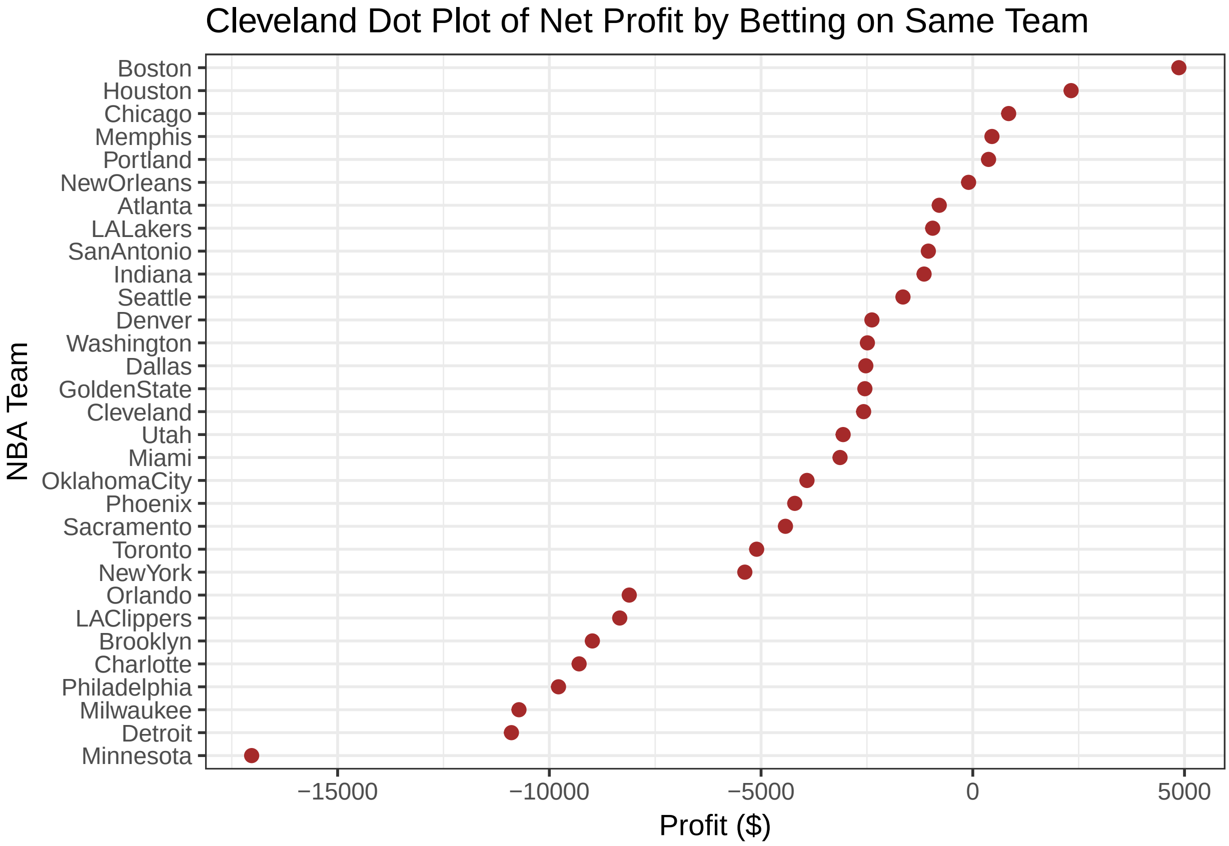
```r
tidy$Payout <- 0
tidy$Payout[1] <- 100
for (i in 2:length(tidy$Payout)) {
  if (tidy$HSpreadChange[i] > 0) {  # if spread change for the game is a hit
    if (tidy$VF[i] - tidy$HF[i] < tidy$HSpreadClose[i]) {    # if home team beat spread
      tidy$Payout[i] <- tidy$Payout[i - 1] + 100
    } else if (tidy$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]) {    # if vis team beat spread
      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")
df <- 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 <- df %>%
  group_by(V) %>% summarise_at("Bet_V_win", sum) %>%
  rename(Team = V)

df_H <- df %>%
  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. Please read our Executive Summary and Conclusion for a succint report of our findings!