GR5702 EDAV Homework 4 Q3

Po-Chieh Liu (pl2441)

```
library(tidyverse)
df <- read_csv('tidy.csv')

hw4_df <- df %>% select(Date,V, H, V1, V2, V3, V4, H1, H2, H3, H4, OU2H) %>%
  mutate(Year = format(Date, "%Y") ) %>%
  mutate(Q3Q4 = V3+V4+H3+H4) %>%
  mutate(Q1Q2 = V1+V2+H1+H2) %>%
  mutate(Q1 = V1+H1) %>%
  select(Year, Q1, Q1Q2, Q3Q4, OU2H, V, H)
```

Quick Introduction

Supreme Court struck down a 1992 federal law in 2018 May to loose the restrictions on sport betting. Later on, there are several on-line betting business services start running. We are interested in how the house gains profits from computing proper money lines, spreads and scores. The dataset we have contains the NBA games' data from season 2007 to 2017 including game date, visiting team, home team, scores of each quarter, house predicted final score, and second half score and other data which are not used in this assignment. The raw data was downloaded from sportbookreviewsonline.com.

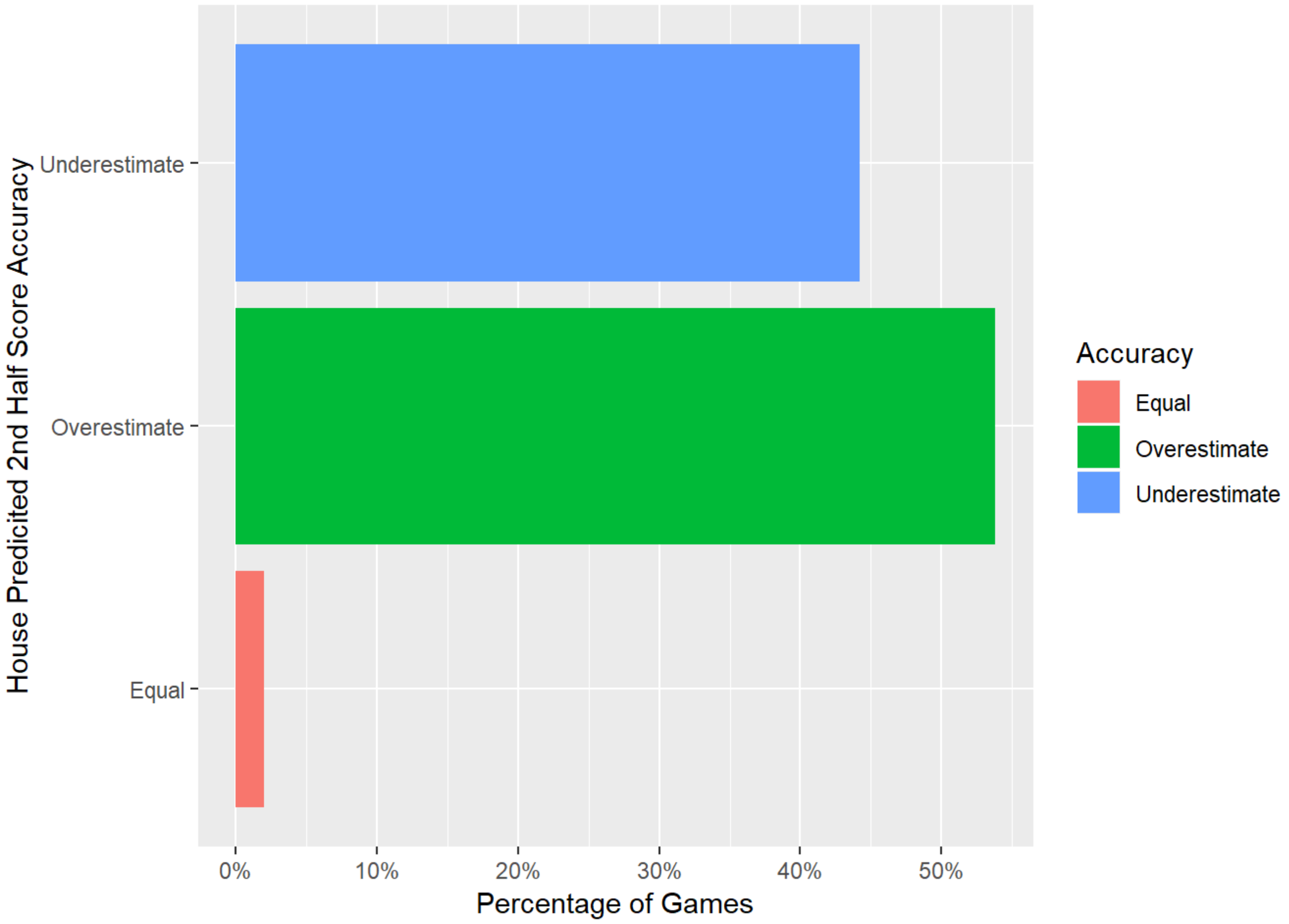
In this assignment, I selected the house estimated 2nd half game score variable for analysis. For betting the 2nd half game score, gambler can either bet the final score is larger than house estimated score, or less than house estimated score. Named bet over and bet under. If the betting result is same as actual game score, then the gambler win. For example, if actual score is smaller than house estimated score and the gambler bet under, then it's a win for the gambler.

As a volleyball player, I think that the 1st quarter (Q1) score and both 1st and 2nd quarter (Q1 and Q2) scores may provide good information for estimating the 2nd half game score. In the following data exploration, I am trying to determine whether the estimated score is a good predictor for competing the house estimated 2nd half score.

House Estimated 2nd Half Score Accuracy

Before start investigating the Q1 and Q1+Q2 predictor performance, we need a baseline model. Following R codes generate a bar chart for comparing the house predicted 2nd half scores versus actual 2nd half scores. From the plot, we can observe that around 54% house estimated scores were larger than actual score. If a gambler always bet under with same amount money, then the gain probability is 53.8%, and lost probability is 46.2% based on historical data.

```
df <- hw4_df %>% select(OU2H, Q3Q4) %>%
  mutate(Accuracy = if_else(OU2H > Q3Q4, 'Overestimate', if_else(OU2H == Q3Q4, 'Equal', 'Unde
restimate')))

ggplot(df, aes(x = Accuracy, fill = Accuracy)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous("Percentage of Games", breaks = c(0, 0.1, 0.2, 0.3, 0.4, 0.5), labels =
 c('0%', '10%', '20%', '30%', '40%', '50%')) +
  scale_x_discrete("House Predicited 2nd Half Score Accuracy") +
  coord_flip()
```



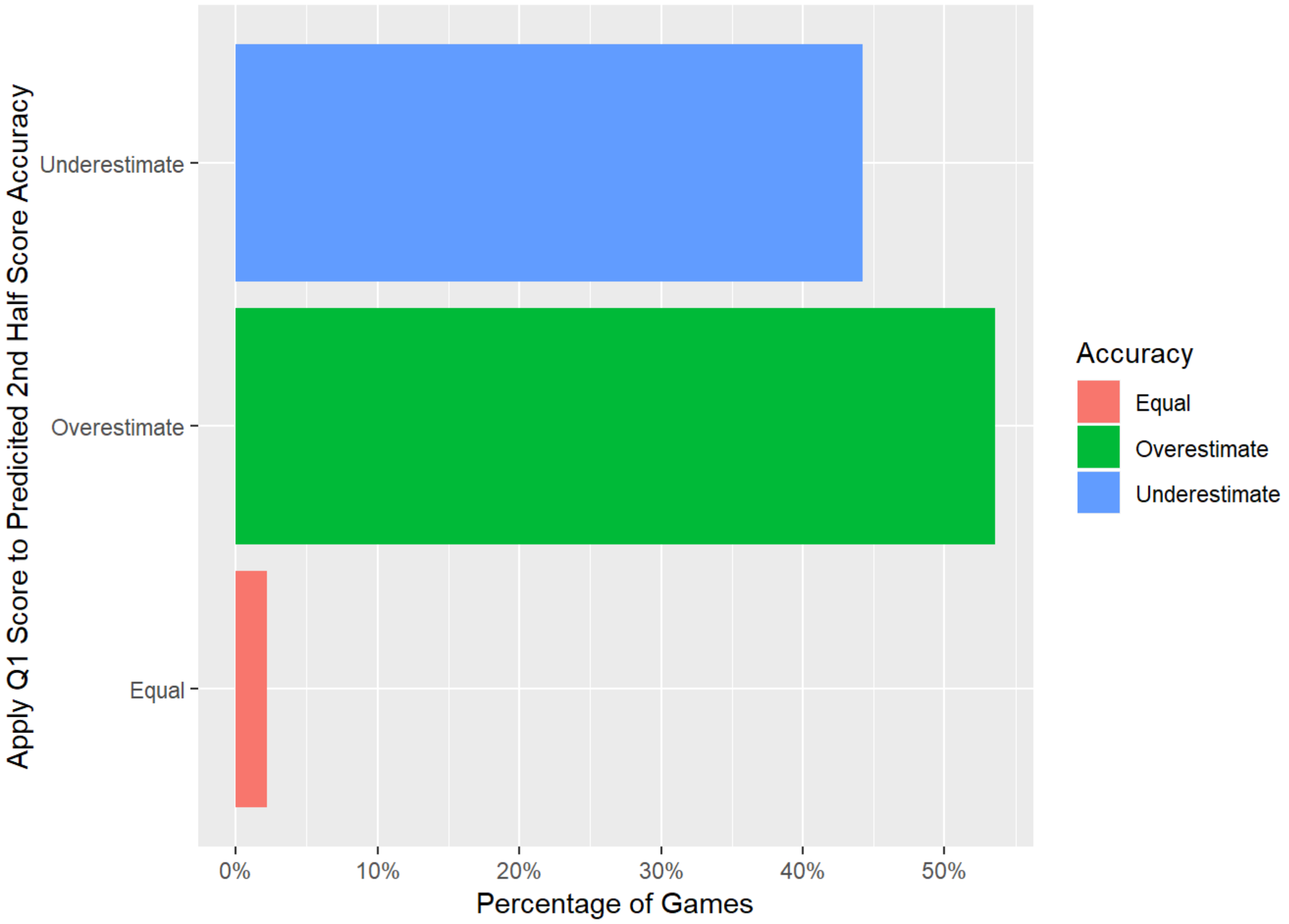
Predicted 2nd Half Score with Q1 Score and Both Q1 and Q2 Score

Although I am a volleyball player instead of a basketball player, I think the first two quarters scores should be good predictors for 2nd half game score. Because the scores provide team some performance information of the game.

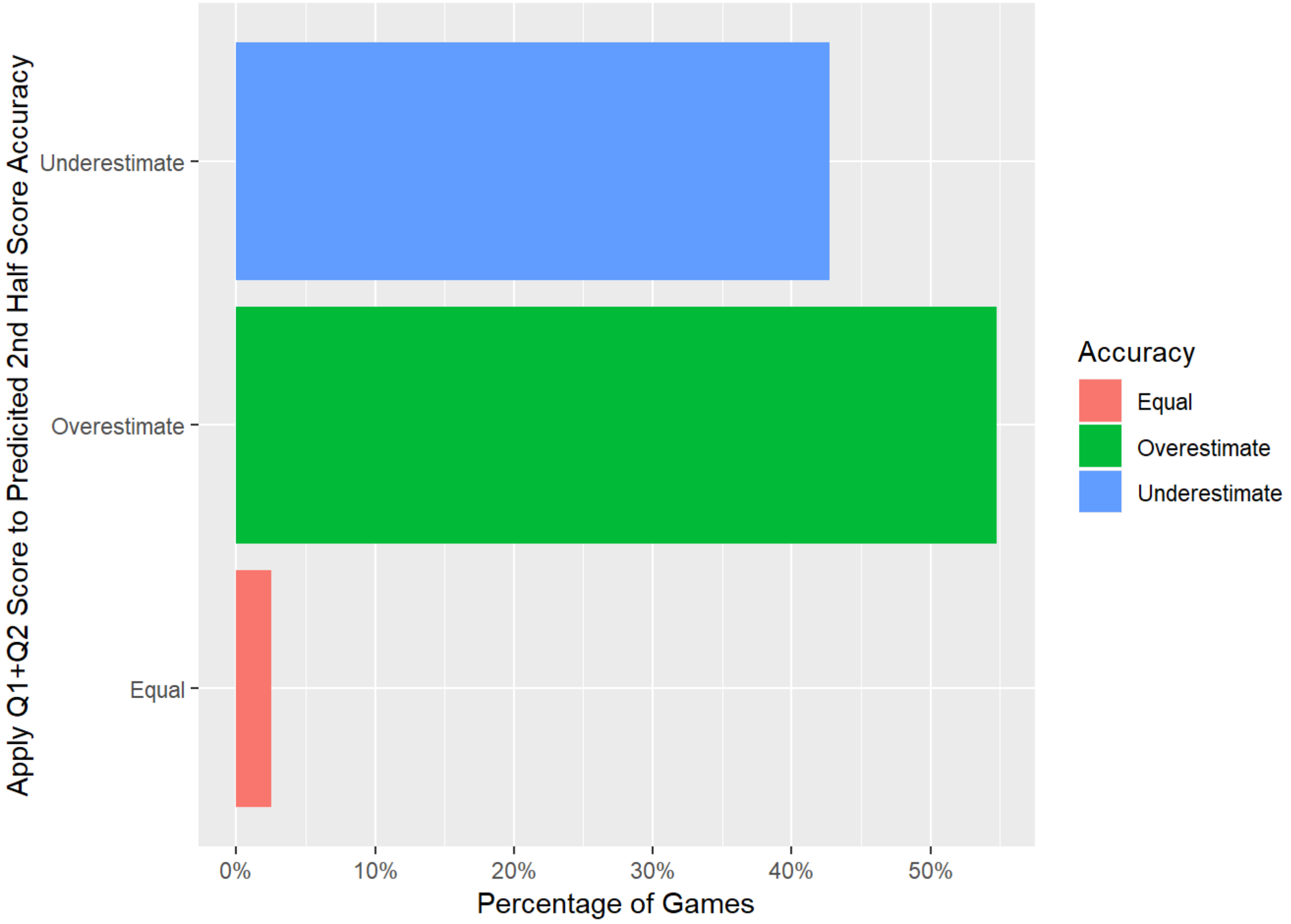
The following R codes generate the predictions accuracy bar charts. For applying Q1 score, I assume both teams perform the same in each quarter. Therefore, the 2nd half score will be 2 times of Q1 score. For applying Q1+Q2 score, I assume both teams perform the same for both 1st and 2nd half games. Therefore, the 2nd half score will be the sum of Q1 and Q2 scores.

From the plots, we can observe that around 53% to 55% game scores were overestimated, and around 45% to 47% games were underestimated. However, the relationship between predicted score and final score does not guarantee bet performance. But if we can improve our prediction accuracy, then we can increase our gain chance, which I plan to explore in the final project work.

```
df <- hw4_df %>% select(Q1, Q3Q4) %>%
  mutate(Accuracy = if_else(Q1*2 == Q3Q4, 'Equal', if_else(Q1*2 > Q3Q4, 'Overestimate', 'Unde
restimate')))

ggplot(df, aes(x = Accuracy, fill = Accuracy)) +
  geom_bar(aes(y=(..count..)/sum(..count..))) +
  scale_y_continuous("Percentage of Games", breaks = c(0, 0.1, 0.2, 0.3, 0.4, 0.5), labels =
 c('0%', '10%', '20%', '30%', '40%', '50%')) +
  scale_x_discrete("Apply Q1 Score to Predicited 2nd Half Score Accuracy") +
  coord_flip()
```



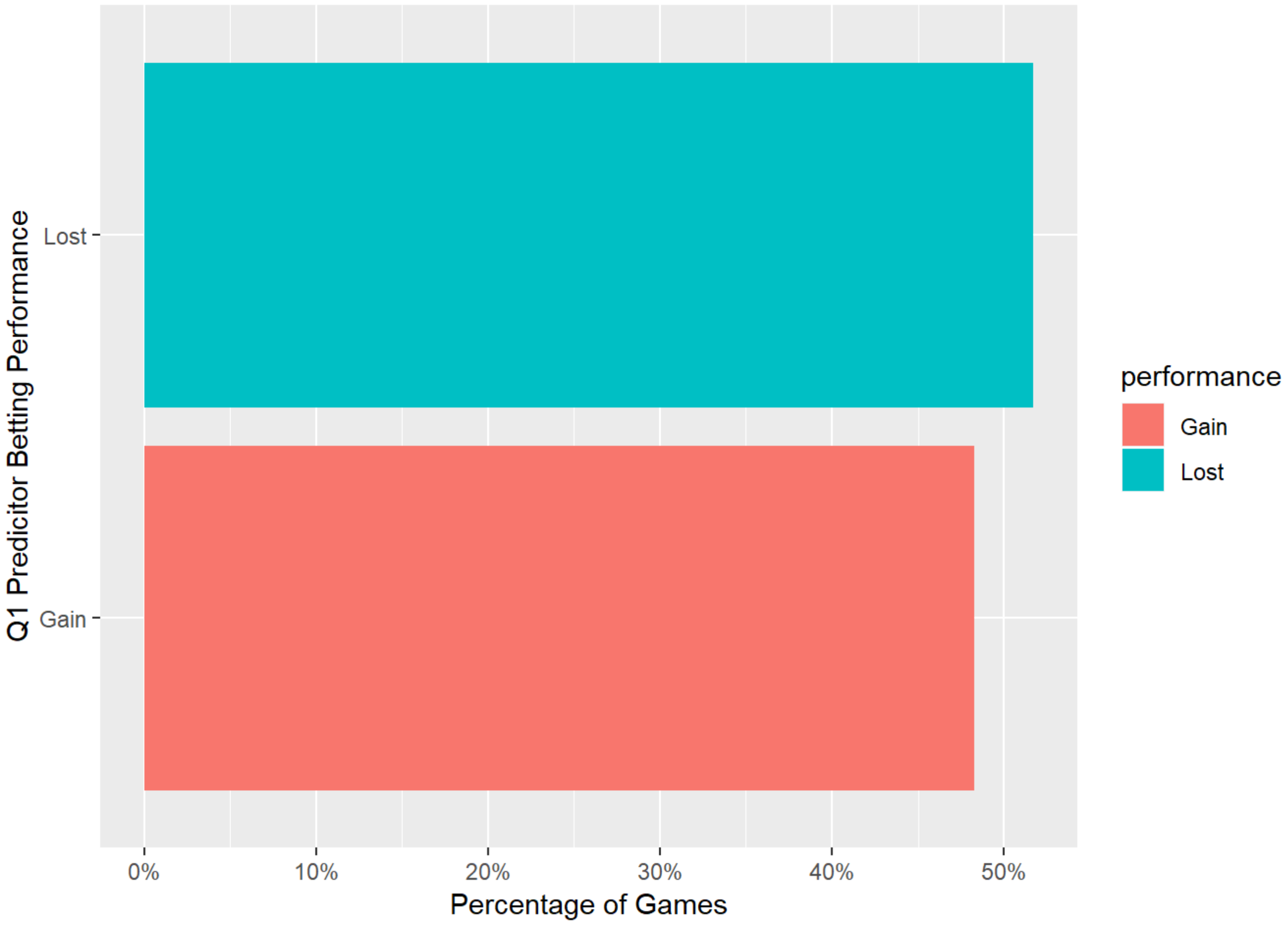
```
df <- hw4_df %>% select(Q1Q2, Q3Q4) %>%
  mutate(Accuracy = if_else(Q1Q2 == Q3Q4, 'Equal', if_else(Q1Q2 > Q3Q4, 'Overestimate', 'Unde
restimate')))

ggplot(df, aes(x = Accuracy, fill = Accuracy)) +
  geom_bar(aes(y=(..count..)/sum(..count..))) +
  scale_y_continuous("Percentage of Games", breaks = c(0, 0.1, 0.2, 0.3, 0.4, 0.5), labels =
 c('0%', '10%', '20%', '30%', '40%', '50%')) +
  scale_x_discrete("Apply Q1+Q2 Score to Predicited 2nd Half Score Accuracy") +
  coord_flip()
```



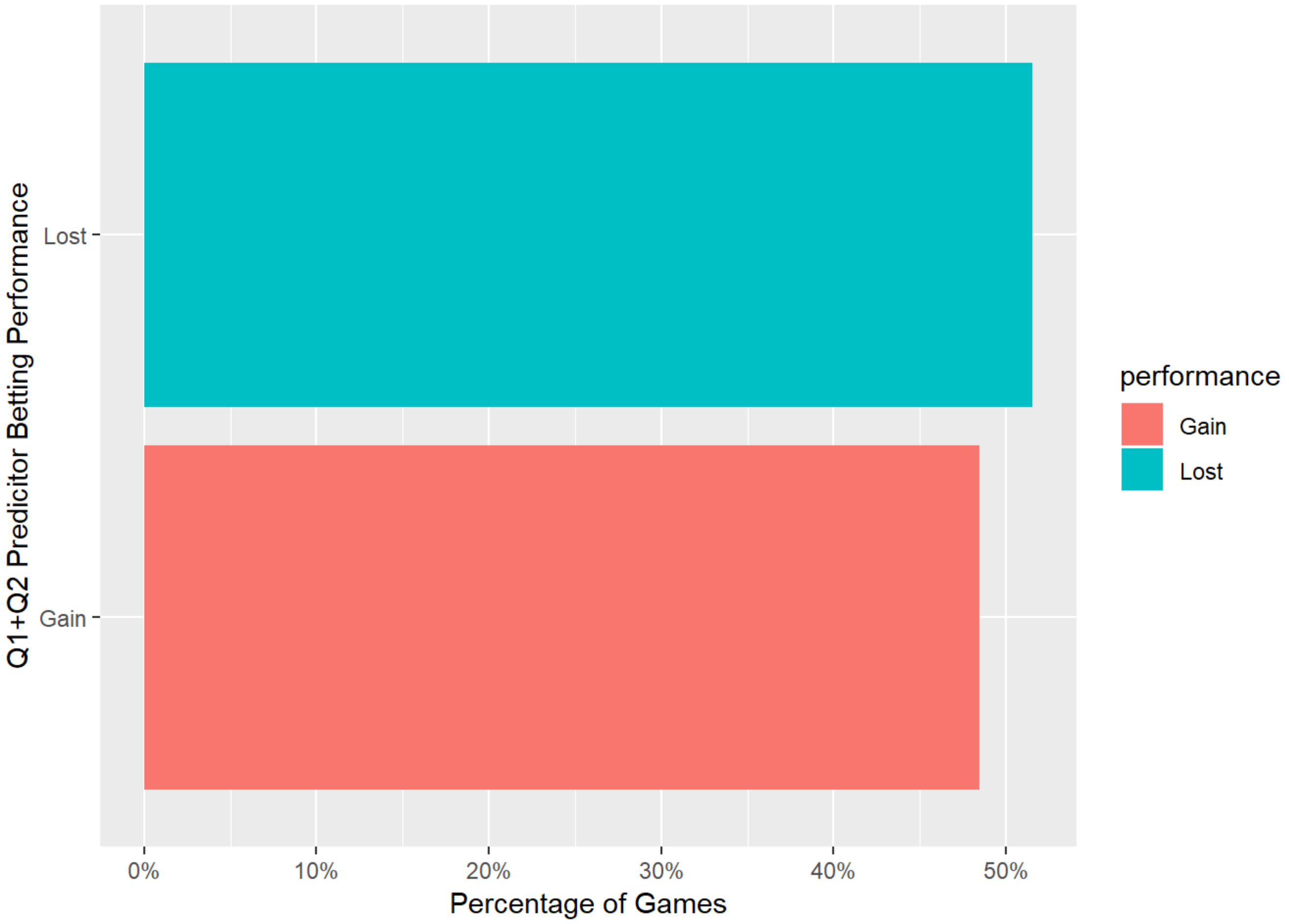
Predicted 2nd Half Score Performance using Q1 score and Both Q1 and Q2 Score

In this section, we tried to predict 2nd half score using the score of first and/or 2nd quarters. Assuming the predicted score is accurate, our betting strategy is to bet over if predicted score is larger than house estimated score, and bet under if predicted score is smaller than house estimated score. The following R codes apply the prediction results to count the gain and lost games, then generate the betting performance plots. From the plots, if we used the prediction results to bet from 2007 to 2017, we will lose around 7,347 (51.7%) games by using Q1 score strategy. With Q1+Q2 strategy, 7,328 (51.6%) games are lost. The overall lose rates are both slighter higher than 50%.

```
df <- hw4_df %>% select(Q1, Q3Q4, OU2H) %>%
  mutate(performance = if_else( ((OU2H>Q3Q4)&(OU2H>2*Q1))|(OU2H<Q3Q4)&(OU2H<2*Q1)|(OU2H==Q3Q
4)&(OU2H==2*Q1) ,'Gain','Lost'))

ggplot(df, aes(x = performance, fill = performance)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous("Percentage of Games", breaks = c(0, 0.1, 0.2, 0.3, 0.4, 0.5), labels =
 c('0%', '10%', '20%', '30%', '40%', '50%')) +
  scale_x_discrete("Q1 Predicitor Betting Performance")+
  coord_flip()
```



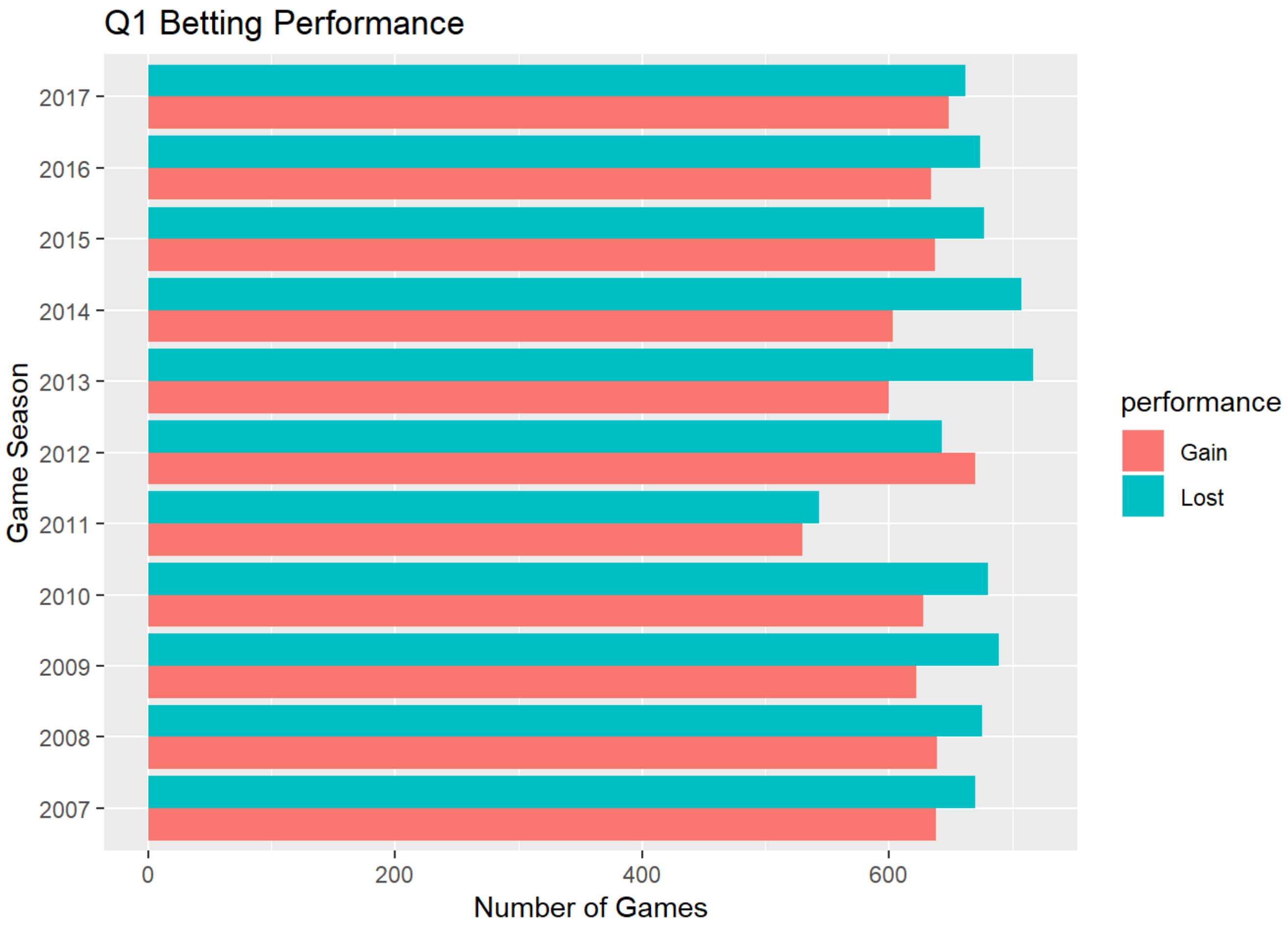
```
df <- hw4_df %>% select(Q1Q2, Q3Q4, OU2H) %>%
  mutate(performance = if_else( ((OU2H>Q3Q4)&(OU2H>Q1Q2))|(OU2H<Q3Q4)&(OU2H<Q1Q2)|(OU2H==Q3Q
4)&(OU2H==Q1Q2) ,'Gain','Lost'))

ggplot(df, aes(x = performance, fill = performance)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  scale_y_continuous("Percentage of Games", breaks = c(0, 0.1, 0.2, 0.3, 0.4, 0.5), labels =
 c('0%', '10%', '20%', '30%', '40%', '50%')) +
  scale_x_discrete("Q1+Q2 Predicitor Betting Performance") +
  coord_flip()
```



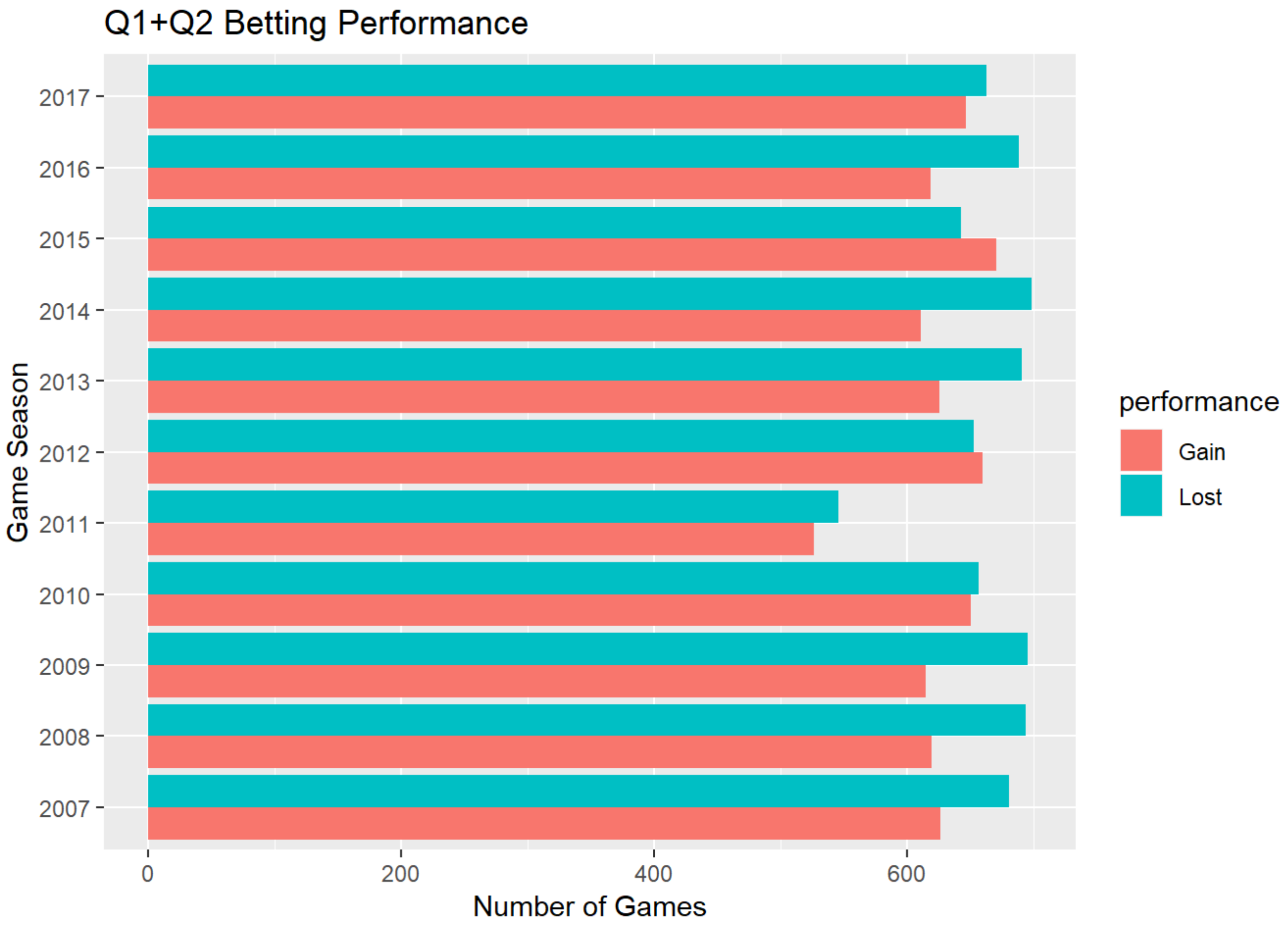
Predicted 2nd Half Score with Q1 and Q2 Score Performance (Yearly)

Previous plots show the betting performances are bad because the gain ratio is below 50%. I am interested to explore the betting performance of each year. Following R codes shows the yearly performances of both predictors. From the plots, we can observe that for using Q1 score, only in 2012 the gain count is greater lost count. For using both Q1 and Q2 scores, only in 2012 and 2015 the gain counts are greater than lost counts.

```
df <- hw4_df %>% select(Q1, Q3Q4, OU2H, Year) %>%
  mutate(performance = if_else( ((OU2H>Q3Q4)&(OU2H>2*Q1))|(OU2H<Q3Q4)&(OU2H<2*Q1)|(OU2H==Q3Q
4)&(OU2H==2*Q1) ,'Gain','Lost'))

ggplot(df, aes(x = Year, fill = performance)) +
  geom_bar(position = "dodge") +
  scale_y_continuous("Number of Games") +
  scale_x_discrete("Game Season") +
  coord_flip() +
  ggtitle("Q1 Betting Performance")
```



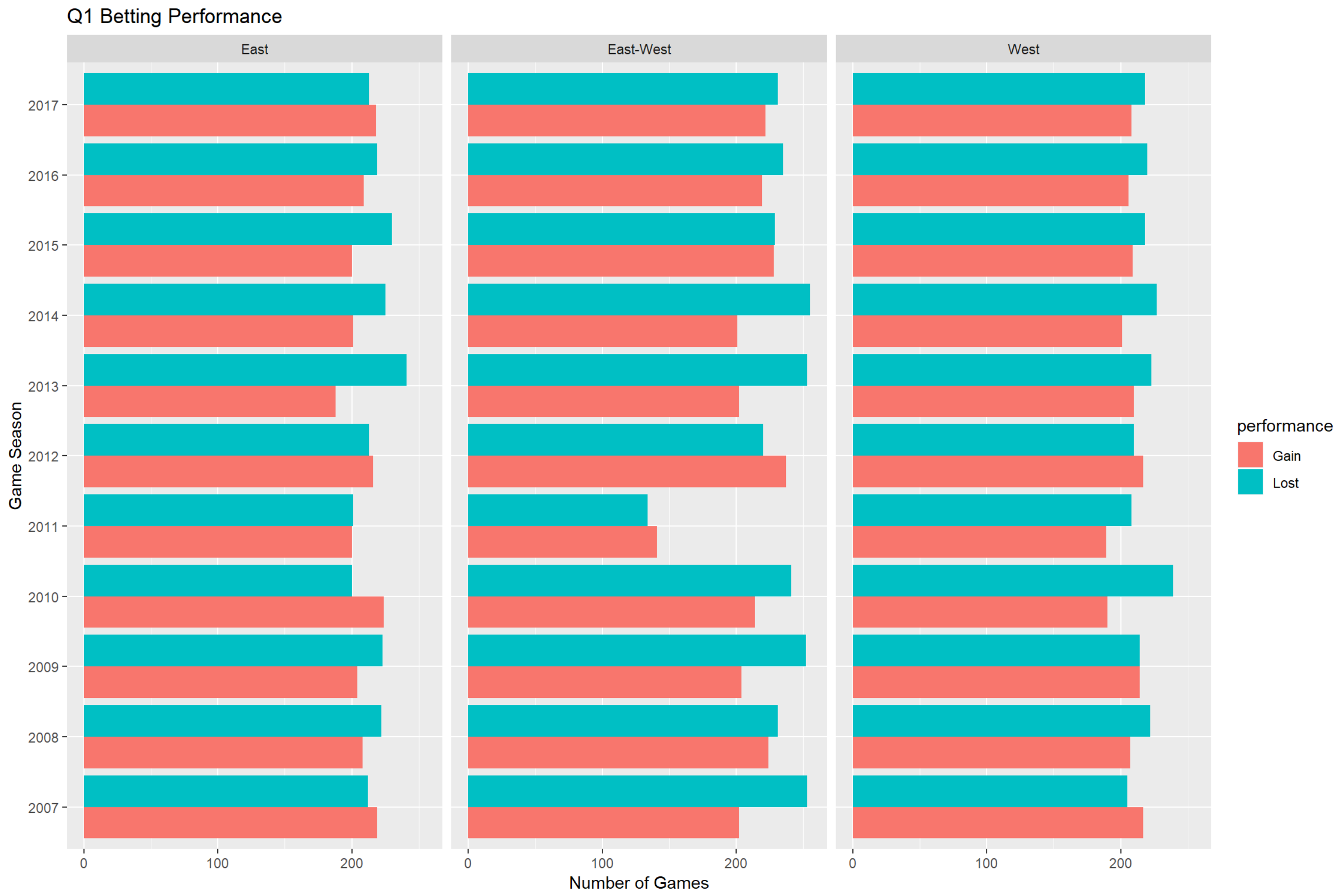
```
df <- hw4_df %>% select(Q1Q2, Q3Q4, OU2H, Year) %>%
  mutate(performance = if_else( ((OU2H>Q3Q4)&(OU2H>Q1Q2))|(OU2H<Q3Q4)&(OU2H<Q1Q2)|(OU2H==Q3Q
4)&(OU2H==Q1Q2) ,'Gain','Lost'))

ggplot(df, aes(x = Year, fill = performance)) +
  geom_bar(position = "dodge") +
  scale_y_continuous("Number of Games") +
  scale_x_discrete("Game Season") +
  coord_flip() +
  ggtitle("Q1+Q2 Betting Performance")
```



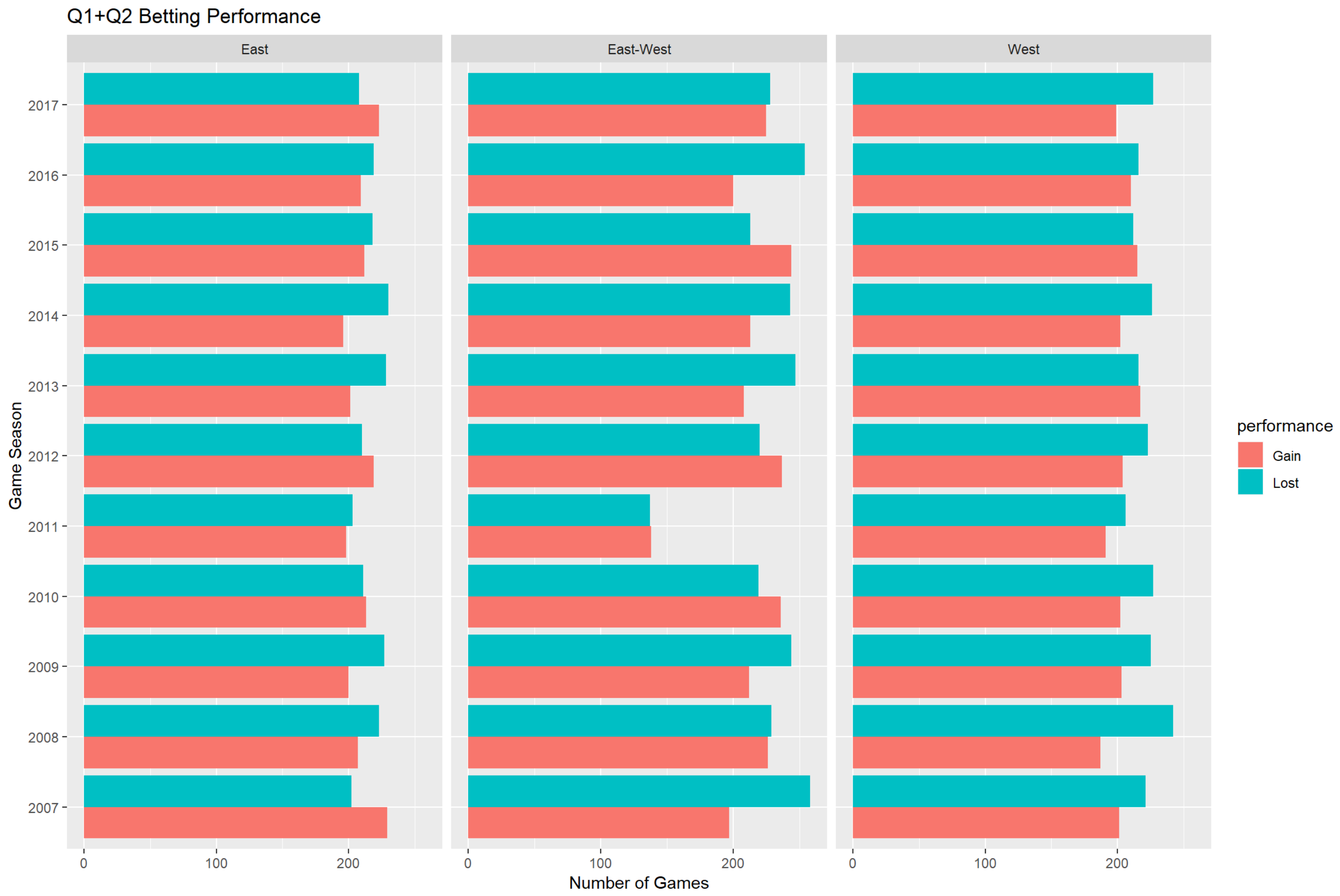
Predicted 2nd Half Score with Q1 and Q2 Score Performance (By Region)

Next, we also explore the performance by grouping the games with teams' region. Q1 score method only have gain counts greater than lost counts in year 2007 and 2012 for west team matches, and in year 2011 and 2012 for east-west team matches, and year 2007, 2010, 2012, 2017 for east team matches. Q1+Q2 score method only have gain counts greater than lost counts in year 2015 for west team matches, and in year 2010, 2011, 2012, 2015 for east-west team matches, and year 2007, 2010, 2012, 2017 for east team matches. The number of years with gain counts greater than lost counts are few. We also can observe that for most year the gaps between lost and gain are huge.

```
East <- c("Atlanta", "Boston", "Brooklyn", "Charlotte", "Chicago", "Cleveland", "Detroit", "I
ndiana", "Miami", "Milwaukee","NewYork","Orlando","Philadelphia","Toronto","Washington")

df <- hw4_df %>% select(V, H, Q1, Q3Q4, OU2H, Year) %>%
  mutate(performance = if_else( ((OU2H>Q3Q4)&(OU2H>2*Q1))|(OU2H<Q3Q4)&(OU2H<2*Q1)|(OU2H==Q3Q
4)&(OU2H==2*Q1) ,'Gain','Lost')) %>%
  mutate(Vside = if_else(V %in% East, 'East', 'West')) %>%
  mutate(Hside = if_else(H %in% East, 'East', 'West')) %>%
  mutate(side = if_else( Vside == Hside, if_else(Vside == 'East', 'East', 'West'), 'East-Wes
t'))

ggplot(df, aes(x = Year, fill = performance)) +
  geom_bar(position = "dodge") +
  scale_y_continuous("Number of Games") +
  scale_x_discrete("Game Season") +
  facet_wrap(~side) +
  coord_flip() +
  ggtitle("Q1 Betting Performance")
```



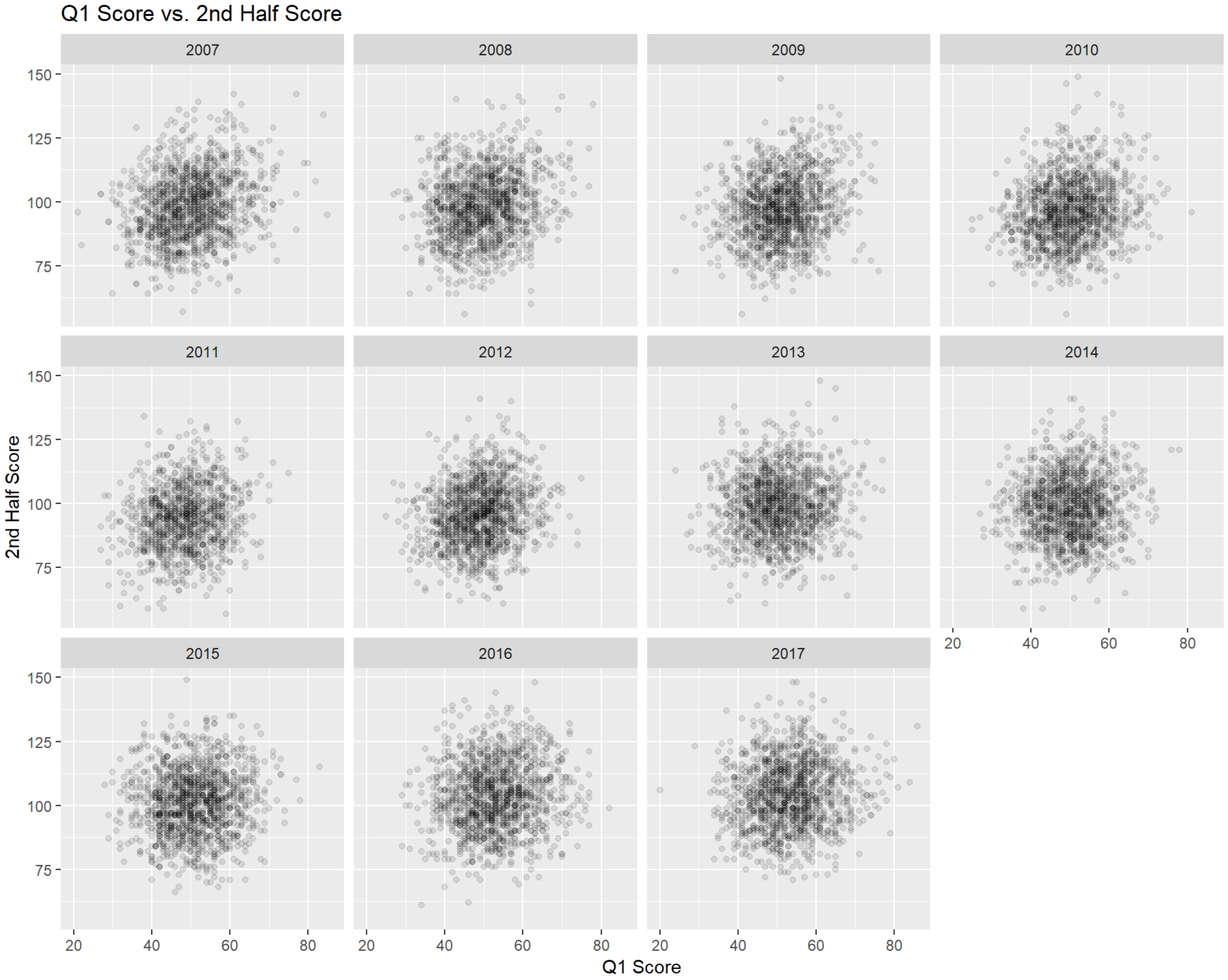
```
df <- hw4_df %>% select(V, H, Q1Q2, Q3Q4, OU2H, Year) %>%
  mutate(performance = if_else( ((OU2H>Q3Q4)&(OU2H>Q1Q2))|(OU2H<Q3Q4)&(OU2H<Q1Q2)|(OU2H==Q3Q
4)&(OU2H==Q1Q2) ,'Gain','Lost')) %>%
  mutate(Vside = if_else(V %in% East, 'East', 'West')) %>%
  mutate(Hside = if_else(H %in% East, 'East', 'West')) %>%
  mutate(side = if_else( Vside == Hside, if_else(Vside == 'East', 'East', 'West'), 'East-Wes
t'))

ggplot(df, aes(x = Year, fill = performance)) +
  geom_bar(position = "dodge") +
  scale_y_continuous("Number of Games") +
  scale_x_discrete("Game Season") +
  facet_wrap(~side) +
  coord_flip() +
  ggtitle("Q1+Q2 Betting Performance")
```



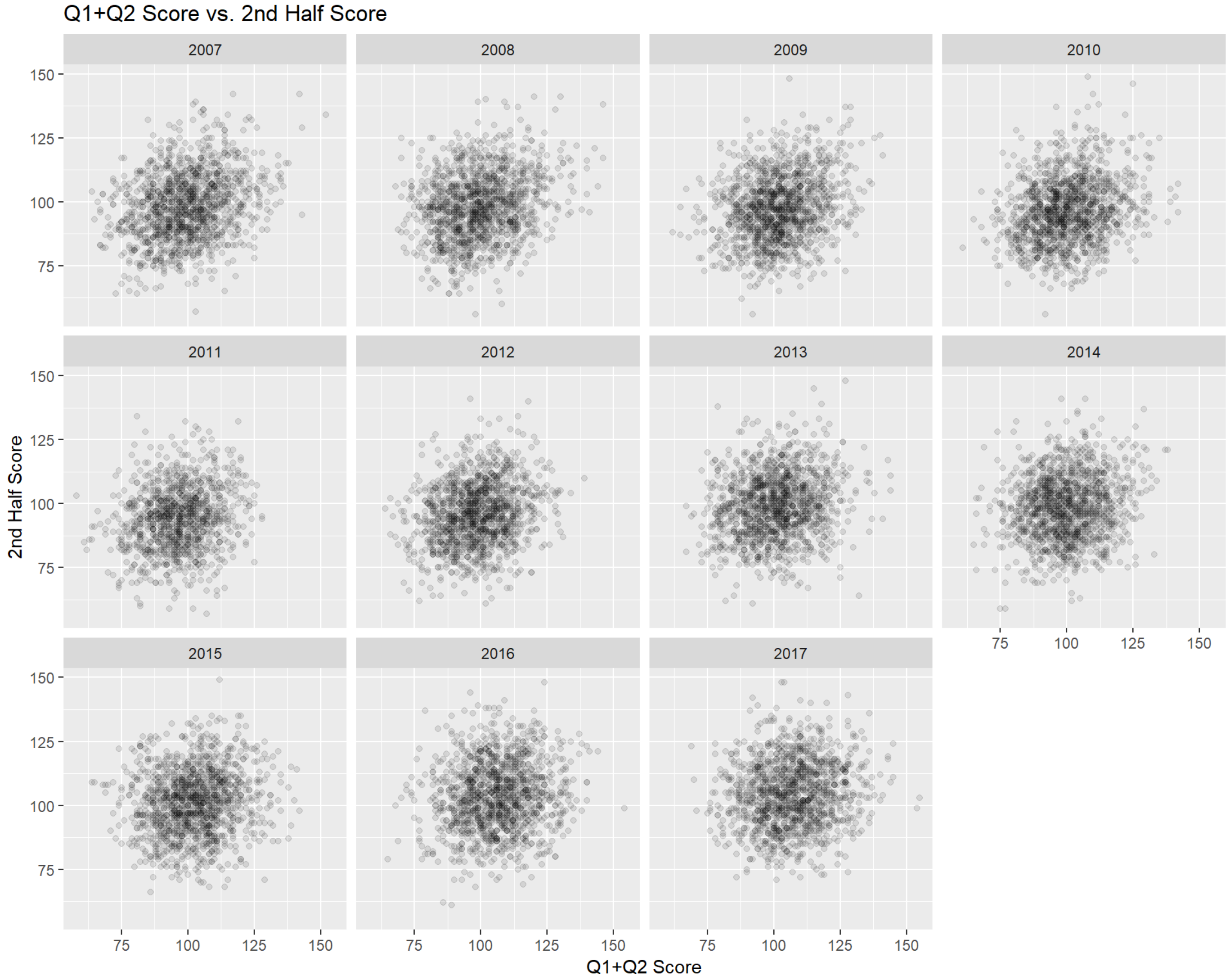
Relationship Between Actual Score and Prediction

Previous plots just used the raw scores to predict the 2nd half game score using a very simple way. The betting performances are bad for both predictors. In order to discover the possible relationship between the predictors and actual scores, the dot plots were generated. From the plots, we can observe the dots show no obvious relationship. All dots are uniformly speeded in the middle of each year plot. No obvious relationship can be visually observed.

```
df <- hw4_df %>% select( Q1, Q1Q2, Q3Q4, Year)

ggplot(df) +
  geom_point(aes(x=Q1,y=Q3Q4), alpha= 0.1) +
  facet_wrap(~Year) +
  scale_x_continuous("Q1 Score") +
  scale_y_continuous("2nd Half Score")+
  ggtitle("Q1 Score vs. 2nd Half Score")
```



```
ggplot(df) +
  geom_point(aes(x=Q1Q2, y=Q3Q4), alpha= 0.1) +
  facet_wrap(~Year) +
  scale_x_continuous("Q1+Q2 Score") +
  scale_y_continuous("2nd Half Score")+
  ggtitle("Q1+Q2 Score vs. 2nd Half Score")
```



Conclusion

The performances of proposed methods in NBA game betting are very bad in compared with baseline model. The gain probabilities are less than 50% for both predictors. However, it makes me more curious about how house design their total score to lure gambler to bet. That score doesn’t have to be accurate, but have to be profitable for house. Moreover, I personally think the Q1 and Q2 scores should provide some information about 2nd half game score. More complex models and other unused variables might increase the accuracy, which will be investigated in the final project.