Cheng-Pang (Tego) Chang

11/05/21

Introduction

You will work your way through this R Markdown document, answering questions as you go along. Please begin l
“author” key in the YAML header. When you’re finished with the document, come back and type your answers int
Please leave all your work below and have your answers where indicated below as well. Please note that we will k
it clear, concise and avoid extremely long printouts. Feel free to add in as many new code chunks as you’d like.

Note:

Throughout this document, the `season` column represents the year each season started. For example, the most
2020-2021, is in the data as season = 2020.

Answers

Question 1: 52.1%
Question 2: 15 out of 16 years.
Question 3: 2013, 674 - 556.
Question 4: WEST: 55.2%, EAST: 48.5%
Question 5: Plotting question, put answer below in the document.
Question 6: Written question, put answer below in the document.
Question 7: 0.32
Question 8: Plotting question, put answer below in the document.
Question 9: Written question, put answer below in the document.
Question 10: Written question with two parts, put answers to (a) and (b) below in the document.
Question 11: Intercept: -20.88, win_pct: 42.40
Question 12: 58.0%
Question 13: The coefficient is -2.875. (Please leave explanation below in the document by answer 13.), EAST: 8
Question 14: Plotting and written question, please put your answers to (a) and (b) below in the document.
Question 15: Written question, put answer below in the document.
Question 16: Written question, put answer below in the document.
Question 17: Written question, put answer below in the document.

```
library(tidyverse)
library(ggplot2)

rm(list = ls())
getwd()
```

```
## [1] "/Users/tegochang/Desktop/Career/US Job Apply/2022Internship/Tech_assessment/sport
```

```
standings <- read_csv("combined_standings.csv")
team_v_team <- read_csv("combined_team_vs_team_records.csv")
```

Conference Standings

Question 1

QUESTION: What is the overall win % to one decimal place (e.g. 41.3%, 74.9%) of teams in the Western Confere
Eastern Conference through the 2005-2020 regular seasons? If the West has more wins than losses against the E
number greater than 50% and vice versa.

```
agg_standings <- aggregate(list(conference_wins = standings$wins),
                           list(conference=standings$conference),
                           sum)
total_wins = agg_standings$conference_wins[agg_standings$conference == "West"] + agg_stan
g_standings$conference == "East"]
Q1_percent <- agg_standings$conference_wins[agg_standings$conference == "West"] / total_w
```

ANSWER 1: 52.1%

Question 2

QUESTION: Out of the past 16 years (2005-2020), how many years has the Western Conference had more wins t

```
agg2_standings <- aggregate(list(conference_wins = standings$wins),
                            list(conference=standings$conference,
                                 season = standings$season),
                            sum)
west_better <- 0
east_better <- 0
diff_west_east_current <- 0
diff_west_east_extreme <- 0
diff_west_east <- c(0)

all_single_season <- data.frame(0, 0)
names(all_single_season)<-c("season","total_season_wins")

for (year in 2005:2020)
{
  single_season <- agg2_standings %>% filter(season == year)
  if (single_season$conference_wins[single_season$conference == 'West'] > single_season$c
ason$conference == 'East'])
  {
    west_better <- west_better + 1
    diff_west_east_current <- single_season$conference_wins[single_season$conference == '
onference_wins[single_season$conference == 'East']
    diff_west_east <- append(diff_west_east, diff_west_east_current)
    if (diff_west_east_extreme < diff_west_east_current)
    {
      diff_west_east_extreme <- diff_west_east_current
    }
  } else{
    east_better <- east_better + 1
    # print(year)
    diff_west_east_current <- single_season$conference_wins[single_season$conference == '
onference_wins[single_season$conference == 'East']
    diff_west_east <- append(diff_west_east, diff_west_east_current)
    if (diff_west_east_extreme < abs(diff_west_east_current))
    {
      diff_west_east_extreme <- abs(diff_west_east_current)
    }
  }

  single_season_agg <- aggregate(list(total_season_wins = single_season$conference_wins),
                                 list(season = single_season$season),
                                 FUN=sum)

  all_single_season <- rbind(all_single_season, single_season_agg)
}

print (west_better); # print (east_better)
```

```
## [1] 15
```

ANSWER 2: 15 out of 16 years.

Question 3

QUESTION: In which year was the disparity between the conferences most extreme? What was the inter confere format WEST WINS - EAST WINS) in that year?

```
all_single_season$diffwins_west_east <- diff_west_east
agg2_standings %>% filter(season == 2013)
```

```
##   conference season conference_wins
## 1       East   2013             556
## 2       West   2013             674
```

ANSWER 3: 2013, 674 - 556.

Question4

QUESTION: For each season, find the playoff team with the lowest win % in each conference. What is the averag
each conference over the 2005-2020 time period? In the case of ties within a season/conference pair, choose just

For example, from the 2020 season, we would include Memphis from the West (38-34, 52.8%) and Washington fr

```
playoffs_no8seed_summary_west <- data.frame()
playoffs_teams <- standings %>% filter(playoffs == "Yes")

for (year in 2005:2020)
{
  playoffs_teams_west <- playoffs_teams %>% filter(season == year & conference == "West"
  playoffs_teams_west_sorted <- playoffs_teams_west[order(playoffs_teams_west$win_pct),]
  playoffs_no8seed_summary_west <- rbind(playoffs_no8seed_summary_west, playoffs_teams_we
}

agg_playoffs_no8seed_summary_west <- aggregate(list(total_wins = playoffs_no8seed_summary
                                                    total_losses = playoffs_no8seed_summa
                                               list(conference = playoffs_no8seed_summary
                                               sum)

total <- agg_playoffs_no8seed_summary_west$total_wins + agg_playoffs_no8seed_summary_west
agg_playoffs_no8seed_summary_west$avg_win_percent <- agg_playoffs_no8seed_summary_west$to

playoffs_no8seed_summary_east <- data.frame()
playoffs_teams <- standings %>% filter(playoffs == "Yes")

for (year in 2005:2020)
{
  playoffs_teams_east <- playoffs_teams %>% filter(season == year & conference == "East"
  playoffs_teams_east_sorted <- playoffs_teams_east[order(playoffs_teams_east$win_pct),]
  playoffs_no8seed_summary_east <- rbind(playoffs_no8seed_summary_east, playoffs_teams_ea
}

agg_playoffs_no8seed_summary_east <- aggregate(list(total_wins = playoffs_no8seed_summary
                                                    total_losses = playoffs_no8seed_summa
                                               list(conference = playoffs_no8seed_summary
                                               sum)

total <- agg_playoffs_no8seed_summary_east$total_wins + agg_playoffs_no8seed_summary_east
agg_playoffs_no8seed_summary_east$avg_win_percent <- agg_playoffs_no8seed_summary_east$to
```

ANSWER 4:

WEST: 55.2% EAST: 48.5%

Question 5

QUESTION: Create a ggplot graph showing the record / win % of playoff and non-playoff teams against playoff a
other conference each season.

For example, your graph should include a visual representation of how Western Conference playoff teams have d
non-playoff teams each season (as well as other combinations of conference and playoffs).

ANSWER 5:

```
all4_conference_against_record <- data.frame(matrix(ncol = 5, nrow = 16))
x <- c("season",
       "west_non-playoff_vs_east_non-playoff",
       "west_playoff_vs_east_non-playoff",
       "west_non-playoff_vs_east_playoff",
       "west_playoff_vs_east_playoff")
colnames(all4_conference_against_record) <- x

conference_against_record <- data.frame(matrix(ncol = 7, nrow = 1))
x <- c("season", "west_playoff", "east_playoff", "west_wins", "west_losses", "west_win_pc
colnames(conference_against_record) <- x

playoffs_combinations <- data.frame(matrix(ncol = 2, nrow = 1))
x <- c("west_playoff", "east_playoff")
colnames(playoffs_combinations) <- x
playoffs_combinations <- rbind(playoffs_combinations, c("No", "No"))
playoffs_combinations <- rbind(playoffs_combinations, c("Yes", "No"))
playoffs_combinations <- rbind(playoffs_combinations, c("No", "Yes"))
playoffs_combinations <- rbind(playoffs_combinations, c("Yes", "Yes"))
playoffs_combinations <- playoffs_combinations[-1,]

for (combination_ind in 1:nrow(playoffs_combinations))
{
  for (year in 2005:2020)
  {
    input <- c(year)
    standings_singleseason <- standings[standings$season == year,]
    # find the non-playoff teams in the west and east as a list
    input <- append(input, c(playoffs_combinations$west_playoff[combination_ind], playoff
off[combination_ind]))
    standings_singleseason_comb_west <- standings_singleseason[standings_singleseason$pla
ations$west_playoff[combination_ind] & standings_singleseason$conference == "West",]
    comb_west_team_singleseason<- standings_singleseason_comb_west$team_short

    standings_singleseason_comb_east <- standings_singleseason[standings_singleseason$pla
ations$east_playoff[combination_ind] & standings_singleseason$conference == "East",]
    comb_east_team_singleseason<- standings_singleseason_comb_east$bb_ref_team_name

    # subset the team_v_team on columns to leave only non-playoff/playoff teams in the we
    team_v_team_singleseason <- team_v_team[team_v_team$season == year,]
    team_v_team_singleseason_comb_west <- team_v_team_singleseason[c("season", "bb_ref_te
_singleseason)]

    # subset the team_v_team on rows to leave only non-playoff/playoff teams in the east
    team_v_team_singleseason_comb_west_comb_east <- team_v_team_singleseason_comb_west[te
omb_west$bb_ref_team_name %in% comb_east_team_singleseason,]

    # filled NA with 0-0 only for our subsetted dataframe
    team_v_team_singleseason_comb_west_comb_east[is.na(team_v_team_singleseason_comb_west

    # parsing the wins and losses based on the strings
    west_losses <- 0
    west_wins <- 0
    col_west_losses <- c()
    col_west_wins <- c()

    for (row_ind in 1:7)
    {
      win_loss_record <- as.numeric(unlist(str_split(team_v_team_singleseason_comb_west_c
],"-")))
      for (ind in 1:length(win_loss_record))
      {
        if (ind %% 2 == 1)
```

```
        {
          west_losses <- west_losses + win_loss_record[ind]
        } else
        {
          west_wins <- west_wins + win_loss_record[ind]
        }
      }
      col_west_losses <- append(col_west_losses, west_losses)
      col_west_wins <- append(col_west_wins, west_wins)
      west_losses <- 0
      west_wins <- 0
    }
    total <- sum(col_west_wins) + sum(col_west_losses)
    input <- append(input,
                    c(sum(col_west_wins), sum(col_west_losses),
                      round(sum(col_west_wins)/total*100, 2),
                      round(sum(col_west_wins)/total*100, 2) - 50
                    ))
    conference_against_record <- rbind(conference_against_record, input)
  }

  conference_against_record <- conference_against_record[-1,]
  conference_against_record$season <- as.numeric(conference_against_record$season)
  conference_against_record$west_playoff <- as.factor(conference_against_record$west_play
  conference_against_record$east_playoff <- as.factor(conference_against_record$east_play
  conference_against_record$west_wins <- as.numeric(conference_against_record$west_wins)
  conference_against_record$west_losses <- as.numeric(conference_against_record$west_loss
  conference_against_record$west_win_pct <- as.numeric(conference_against_record$west_win
  conference_against_record$west_win_more_pct <- as.numeric(conference_against_record$wes

  # store records of all combinations
  all4_conference_against_record$season <- conference_against_record$season
  if (combination_ind == 1){
    all4_conference_against_record$`west_non-playoff_vs_east_non-playoff` <- conference_a
ct
  } else if(combination_ind == 2){
    all4_conference_against_record$`west_playoff_vs_east_non-playoff` <- conference_again
  } else if(combination_ind == 3){
    all4_conference_against_record$`west_non-playoff_vs_east_playoff` <- conference_again
  } else{
    all4_conference_against_record$`west_playoff_vs_east_playoff` <- conference_against_r
  }
  conference_against_record <- data.frame(matrix(ncol = 7, nrow = 1))
  x <- c("season", "west_playoff", "east_playoff", "west_wins", "west_losses", "west_win_
")
  colnames(conference_against_record) <- x
}
mean(all4_conference_against_record$`west_non-playoff_vs_east_non-playoff`)
```

```
## [1] 55.50312
```

```
mean(all4_conference_against_record$`west_playoff_vs_east_non-playoff`)
```

```
## [1] 78.1875
```

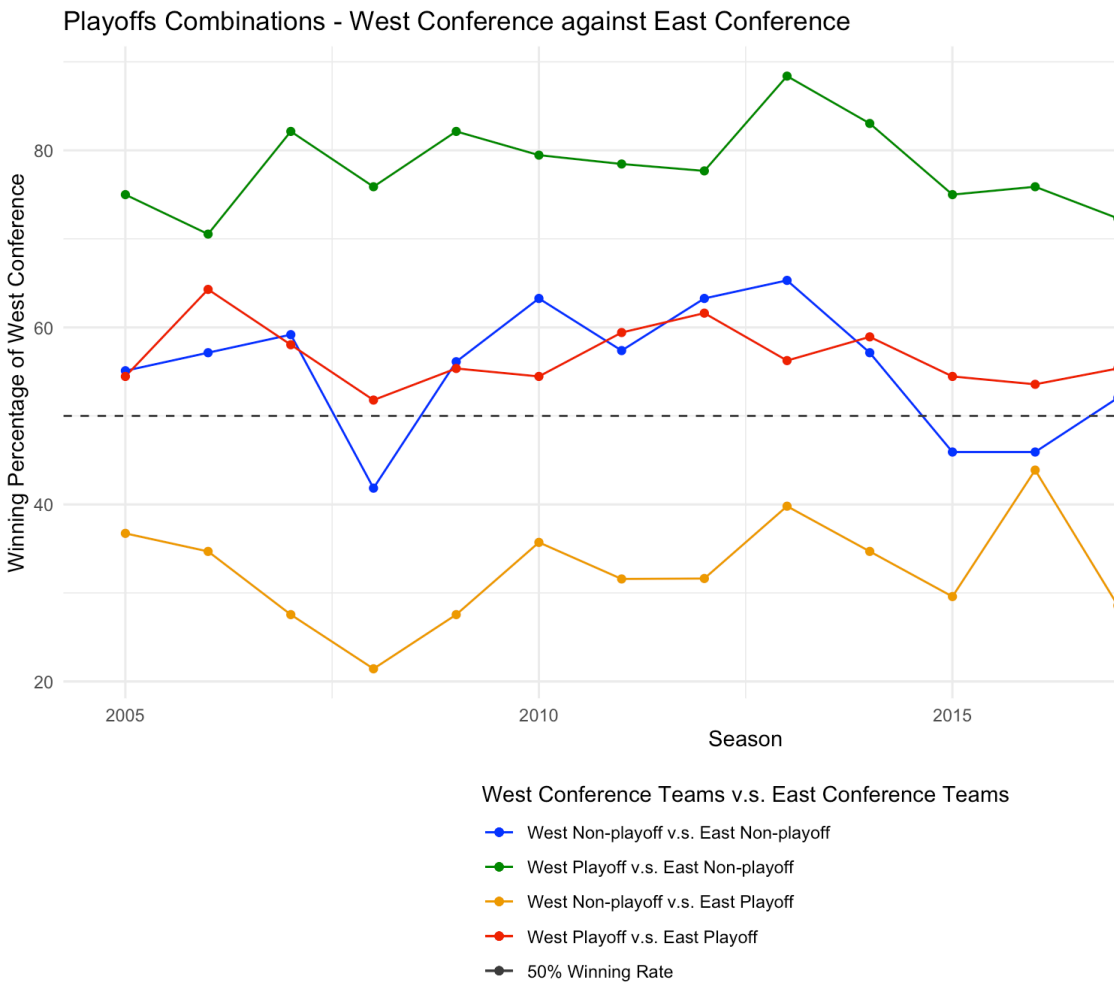
```
mean(all4_conference_against_record$`west_non-playoff_vs_east_playoff`)
```

```
## [1] 31.9425
```

```
mean(all4_conference_against_record$`west_playoff_vs_east_playoff`)
```

```
## [1] 57.24312
```

```
ggplot(all4_conference_against_record, aes(x=season)) +
  geom_line(aes(y = `west_non-playoff_vs_east_non-playoff`, colour = "West Non-playoff v.
  geom_line(aes(y = `west_playoff_vs_east_non-playoff`, colour = "West Playoff v.s. East
  geom_line(aes(y = `west_non-playoff_vs_east_playoff`, colour = "West Non-playoff v.s. E
  geom_line(aes(y = `west_playoff_vs_east_playoff`, colour = "West Playoff v.s. East Play
  geom_hline(aes(yintercept=50, colour="50% Winning Rate"), linetype="dashed") +
  geom_point(aes(x=season,y=`west_non-playoff_vs_east_non-playoff`, colour = "West Non-pl
off")) +
  geom_point(aes(x=season,y=`west_playoff_vs_east_non-playoff`, colour = "West Playoff v.
  geom_point(aes(x=season,y=`west_non-playoff_vs_east_playoff`, colour = "West Non-playof
  geom_point(aes(x=season,y=`west_playoff_vs_east_playoff`, colour = "West Playoff v.s. E
  labs(title="Playoffs Combinations - West Conference against East Conference",
       x="Season",
       y="Winning Percentage of West Conference") +
  scale_color_manual(name = "West Conference Teams v.s. East Conference Teams",
                     values = c("West Non-playoff v.s. East Non-playoff" = "blue1",
                                "West Playoff v.s. East Non-playoff" = "green4",
                                "West Non-playoff v.s. East Playoff" = "orange2",
                                "West Playoff v.s. East Playoff" = "red2",
                                "50% Winning Rate" = "grey25"
                     )) +
  theme_minimal() +
  theme(legend.position = "bottom", legend.direction = "vertical")
```



Question 6

QUESTION: Write up to two sentences describing any takeaways you have from your visual above that you coul data science department.

ANSWER 6:

In the past 16 seasons, considering the records among playoff teams, represented by the red line in the figure ab playoff teams, represented by the blue line, teams in the west conference have a better chance of winning agains conference by 7.2% and 5.5% correspondingly. On the other hand, considering the other two combinations of pl teams, represented by the green and orange lines, the ones in the west conference still outperform the ones in th calculated by 78.2% - (100% - 31.9%).

Point Margins and Schedules

Question 7

In this next series of questions we are going to look at the strength of schedule by examining teams' opponents'

First, calculate the average point margin for each team's opponents in each year, weighting by the number of time example, if team A played against team B once and team C twice, and team B had a season average point margi season average point margin of -3, team A's opponents would have an average point margin of (2 - 3 - 3) / 3 = -1. the results of this calculation here, you are going to use it for the next few questions.

```
standings$point_margins <- standings$points_scored_per_game - standings$points_allowed_pe
team_v_team_okc <- team_v_team[,c("season", "bb_ref_team_name", "OKC")]
team_v_team_okc_2016 <- team_v_team_okc %>% filter(season == 2016)

for (row_ind in 1:30)
{
  team_v_team_okc_2016$games[row_ind] <- sum(as.numeric(unlist(str_split(team_v_team_okc_
)))
}

for (team_name in team_v_team_okc_2016$bb_ref_team_name)
{
  team_v_team_okc_2016$point_margins[team_v_team_okc_2016$bb_ref_team_name == team_name]
ns[standings$bb_ref_team_name == team_name & standings$season == 2016]
}
team_v_team_okc_2016$games[is.na(team_v_team_okc_2016$games)] <- 0
# sum(team_v_team_okc_2016$games)
team_v_team_okc_2016$total_point_margins <- team_v_team_okc_2016$games * team_v_team_okc_
```

QUESTION: What was OKC's opponents' average point margin (to two decimal places) in the 2016 season?

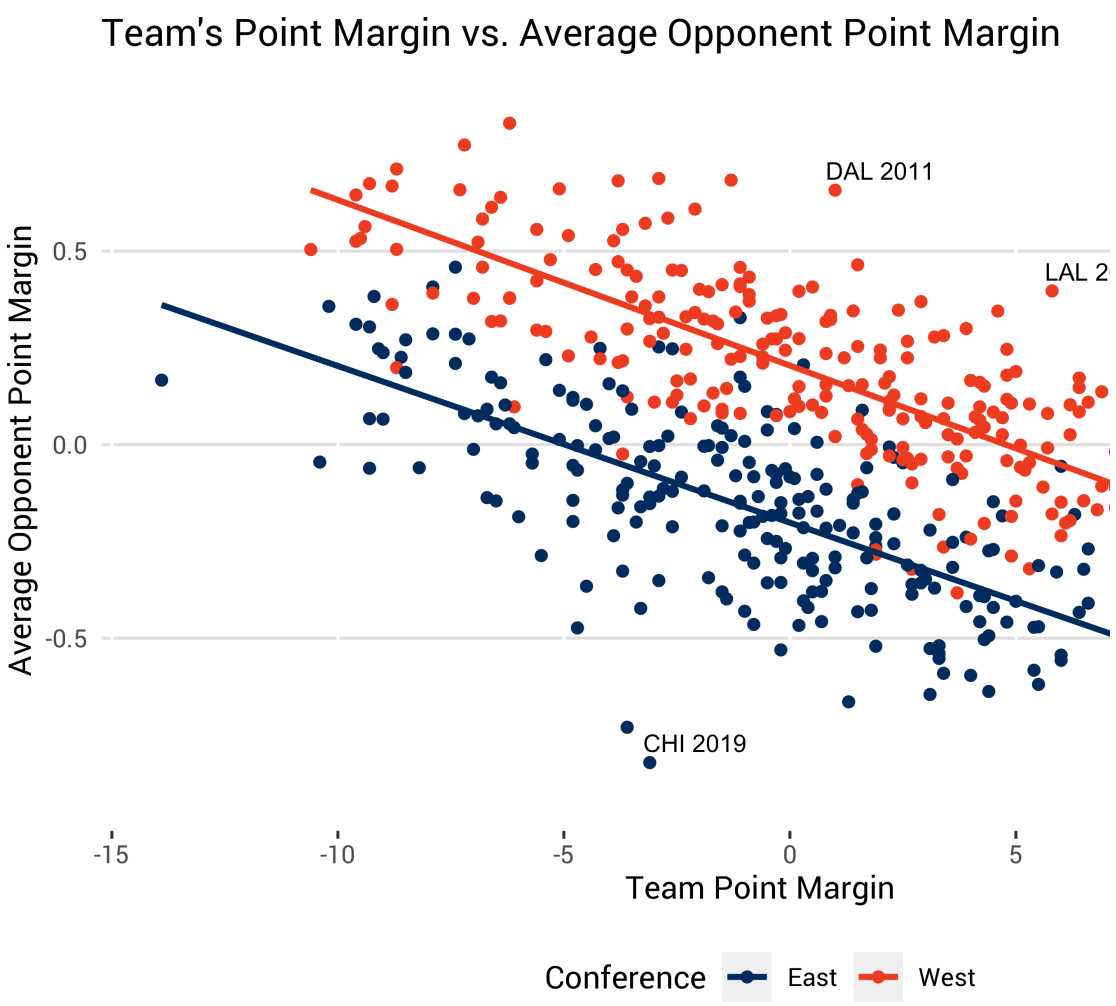
```
avg_point_margins <- sum(team_v_team_okc_2016$total_point_margins) / 82
```

ANSWER 7: 0.32

Question 8

QUESTION: As close as you can, reproduce the following plot. There is one point on the plot for each team/seas

ANSWER 8:

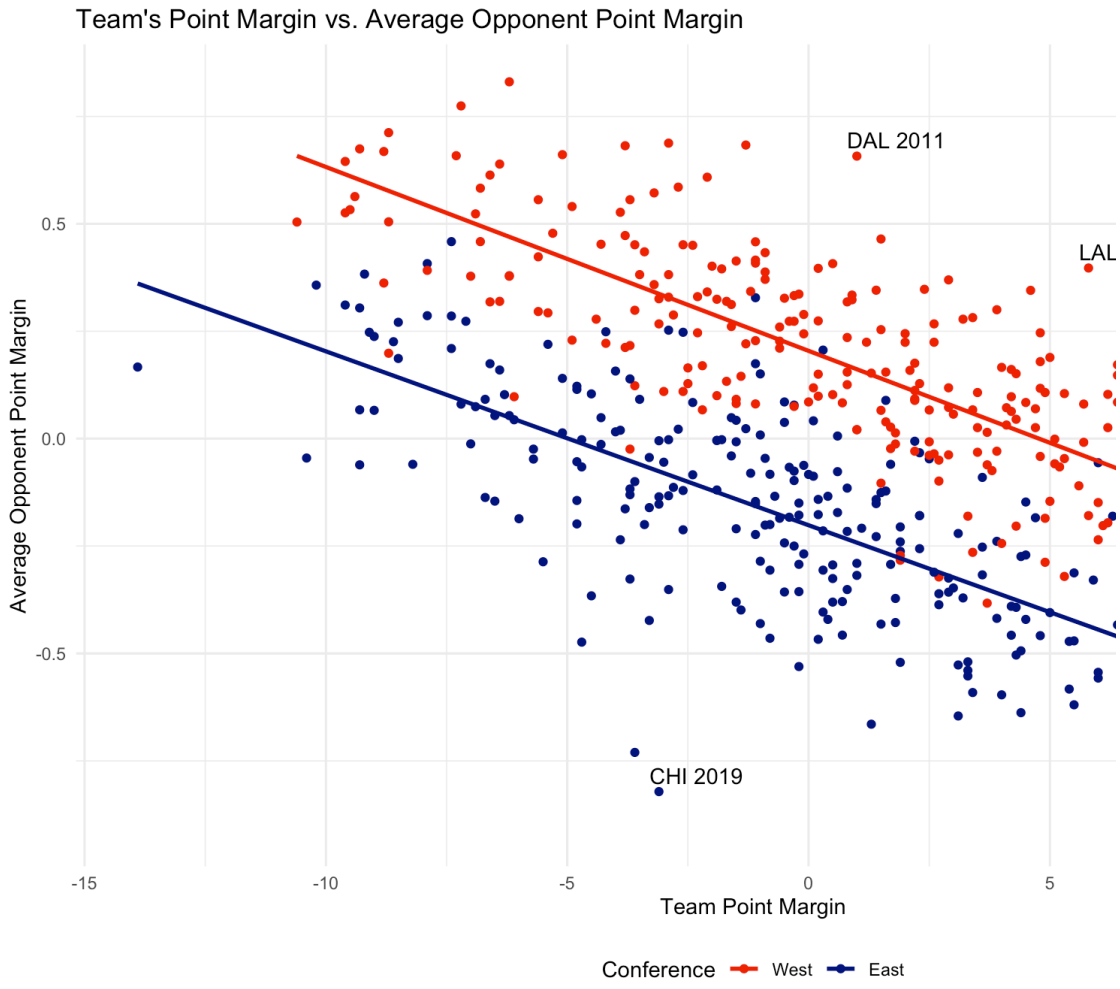

Team's Point Margin vs. Average Opponent Point Margin
DAL 2011
LAL 2
CHI 2019
Average Opponent Point Margin
0.5
0.0
-0.5
-15
-10
-5
0
5
Team Point Margin
Conference
East
West

```
strengh_schedule <- data.frame(matrix(ncol = 5, nrow = 1))
x <- c("season", "conference", "team_short", "avg_opponent_point_margins", "team_point_ma
colnames(strengh_schedule) <- x

for (year in 2005:2020)
{
  team_v_team_single_year <- team_v_team %>% filter(season == year)
  all_team_short <- standings$team_short[standings$season == year]

  for (team_short in all_team_short)
  {
    input <- c(year)
    input <- append(input, c(standings$conference[standings$team_short == team_short & st
team_short))
    team_v_team_single_year_single_team <- team_v_team_single_year[,c("season", "bb_ref_t

    # get total games played with each team in a season
    for (row_ind in 1:30)
    {
      team_v_team_single_year_single_team$games[row_ind] <- sum(as.numeric(unlist(str_spl
ar_single_team[row_ind,team_short],"-"))))
    }

    for (team_long in team_v_team_single_year$bb_ref_team_name)
    {
      team_v_team_single_year_single_team$point_margins[team_v_team_single_year_single_te
eam_long] <- standings$point_margins[standings$bb_ref_team_name == team_long & standings$
    }

    # Team can not play against itself, so fill with 0
    team_v_team_single_year_single_team$games[is.na(team_v_team_single_year_single_team$g
    sum(team_v_team_single_year_single_team$games)
    team_v_team_single_year_single_team$total_point_margins <- team_v_team_single_year_si
_team_single_year_single_team$point_margins
    avg_point_margins <- sum(team_v_team_single_year_single_team$total_point_margins) / s
ar_single_team$games)
    input <- append(input, c(avg_point_margins, standings$point_margins[standings$team_sh
dings$season == year]))
    strengh_schedule <- rbind(strengh_schedule, input)
    input <- c()
  }
}
strengh_schedule <- strengh_schedule[-1,]
str(strengh_schedule)
```

```
## 'data.frame':    480 obs. of  5 variables:
##  $ season                   : chr  "2005" "2005" "2005" "2005" ...
##  $ conference               : chr  "East" "East" "East" "East" ...
##  $ team_short               : chr  "ATL" "BOS" "BKN" "CHA" ...
##  $ avg_opponent_point_margins: chr  "0.121951219512196" "-0.00731707317073112" "-0.228
73170732" ...
##  $ team_point_margins       : chr  "-4.8" "-1.5" "1.39999999999999" "-4" ...
```

```
strengh_schedule$season <- as.numeric(strengh_schedule$season)
strengh_schedule$conference <- as.factor(strengh_schedule$conference)
strengh_schedule$avg_opponent_point_margins <- as.numeric(strengh_schedule$avg_opponent_p
strengh_schedule$team_point_margins <- as.numeric(strengh_schedule$team_point_margins)

label_data <- strengh_schedule[strengh_schedule$season == 2019 & strengh_schedule$team_sh
                               strengh_schedule$season == 2019 & strengh_schedule$team_sh
                               strengh_schedule$season == 2011 & strengh_schedule$team_sh

label_data$name <- c("DAL 2011", "CHI 2019", "LAL 2019")

ggplot(strengh_schedule, aes(x=team_point_margins,
                             y=avg_opponent_point_margins,
                             color=conference)) +
  geom_point() +
  geom_text(data=label_data,vjust = -0.5 , hjust = 0.1, color = "black",
            aes(team_point_margins,avg_opponent_point_margins,label=name)) +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title="Team's Point Margin vs. Average Opponent Point Margin",
             x="Team Point Margin",
             y="Average Opponent Point Margin") +
  scale_color_manual(name = "Conference",
                     values = c("West" = "red2", "East" = "navyblue")) +
  theme_minimal() +
  theme(legend.position = "bottom")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Question 9

QUESTION: Write no more than 4 sentences explaining this plot and two takeaways to a non-technical member

ANSWER 9: Each point in the figure represents each team in the seasons from 2005 to 2020. The more close the the more points the team (that the point stands for) scores than the points it allows, as the x-axis, in a more stron the y-axis. Given the same strength level of schedule, teams in the west conference generally win by more points east conference. The two points representing Los Angeles Lakers in 2019 and Dallas Mavericks in 2011 are very which stands for they are both very competitive teams, and they both won NBA championships in the correspond

Question 10

PART (a): What do you conclude about the relative strength of schedule for the two labeled teams (DAL 2011 and rest of the teams) at the top of the plot? Please answer in 1 sentence.

ANSWER 10 (a): Given the same strength level of schedule, meaning the same value in the y-axis, DAL 2011 and best teams in outscoring their opponents.

PART (b): Do you have any hypotheses as to why teams from 2019 and 2011 appear at the extremes of this grap than 3 sentences.

ANSWER 10 (b): As the number of total games played, the sample size, reduced in these two seasons, the extrer likely to happen. Teams in the 2011 season play fewer games, reducing from 82 to 66 games, due to the NBA loc due to COVID-19, only selected teams keep playing games in Bubble.

Logistic Regression

Question 11

QUESTION: Fit a logistic regression model on all of the data 2005-2020 predicting the chance a team makes the that season. What are the coefficients?

```
standings$playoffs_num <- 0
standings$playoffs_num[standings$playoffs == "Yes"] <- 1

model1 <- glm(playoffs_num ~ win_pct, data = standings, family = binomial(link = logit))
# summary(model1)

# the below model might be more appropriate for interpretation
standings$win_pct100 <- standings$win_pct*100
model2 <- glm(playoffs_num ~ win_pct100, data = standings, family = binomial(link = logit
# summary(model2)
```

ANSWER 11: Intercept: -20.88, win_pct: 42.40

Question 12

QUESTION: Using your model from the previous question, what is the probability that a team with exactly a 50% (rounded to the nearest decimal, e.g. 44.7%)?

```
test_data <- standings[standings$win_pct == 0.5,][1,]
predict(model1, test_data,type='response')*100
```

```
##        1
## 57.95795
```

ANSWER 12: 58.0%.

Question 13

Add a indicator variable called `is_west` to your regression model that is TRUE if the team is in the Western Con

```
standings$is_west <- FALSE
standings$is_west[standings$conference == "West"] <- TRUE
# sum(standings$is_west)
```

QUESTION: What is the `is_west` coefficient and what does it mean? What is the prediction to the nearest decir team in the West with a 50% win rate?

```
model3 <- glm(playoffs_num ~ win_pct + is_west, data = standings, family = binomial(link
summary(model3)
```

```
##
## Call:
## glm(formula = playoffs_num ~ win_pct + is_west, family = binomial(link = logit),
##     data = standings)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.60871  -0.04162   0.00096   0.09565   2.36506
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -25.6315     3.2472  -7.893 2.94e-15 ***
## win_pct      54.4880     6.8085   8.003 1.22e-15 ***
## is_westTRUE  -2.8751     0.5751  -4.999 5.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 663.29  on 479  degrees of freedom
## Residual deviance: 127.34  on 477  degrees of freedom
## AIC: 133.34
##
## Number of Fisher Scoring iterations: 8
```

```
# to interpret the meaning of coeficient of the is_west predictor, we shall center the wi
standings$win_pct100_centered <- standings$win_pct100 - mean(standings$win_pct100)
model_inter <- glm(playoffs_num ~ win_pct100_centered + is_west, data = standings, family
))
summary(model_inter)
```

```
##
## Call:
## glm(formula = playoffs_num ~ win_pct100_centered + is_west, family = binomial(link = l
##     data = standings)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.60871  -0.04162   0.00096   0.09565   2.36506
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.59525    0.35011   4.556 5.20e-06 ***
## win_pct100_centered  0.54488    0.06809   8.003 1.22e-15 ***
## is_westTRUE         -2.87514    0.57509  -4.999 5.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 663.29  on 479  degrees of freedom
## Residual deviance: 127.34  on 477  degrees of freedom
## AIC: 133.34
##
## Number of Fisher Scoring iterations: 8
```

```
# the probability to make the playoffs for east conference team (is_west = 0)
logit_ori <- exp(1.60)*exp(0.54)
prob_ori <- logit_ori / (1 + logit_ori) *100

# the probability to make the playoffs for west conference team (is_west = 1)
logit <- exp(1.60)*exp(0.54)*exp(-2.88)
prob <- logit / (1 + logit) * 100


test_data_2 <- standings[standings$win_pct == 0.5,][c(1,5),]
predict(model3, test_data_2,type='response')*100
```

```
##        1        2
## 83.37586 22.05200
```

ANSWER 13:
The coefficient is -2.875. Explanation: considering a team in the east conference with the average winning percen team makes playoffs is 89.5%. However, when the team is in the west conference with the same winning percent team makes playoffs will drop 57.2% to 32.3%. That is the effect of the coefficient of is_west in this model based

EAST: 83.4%
WEST: 22.1%

Question 14

We are going to investigate whether it's possible that the relationship you found in the previous question could be randomness. We're only looking at 30 teams over 16 years, so sample size might be a concern. To do this, you w

For each of 10,000 iterations, randomly reorder the conference labels so that in each iteration, there are 15 rando teams labeled as West. For example, in a given iteration, we might have assigned OKC to the East and BKN to th assigned to East. For each iteration, fit a new logistic regression model with the same variables as in question 13 from win % and `is_west`) and extract the `is_west` coefficient. Save all 10,000 `is_west` coefficients in a vect

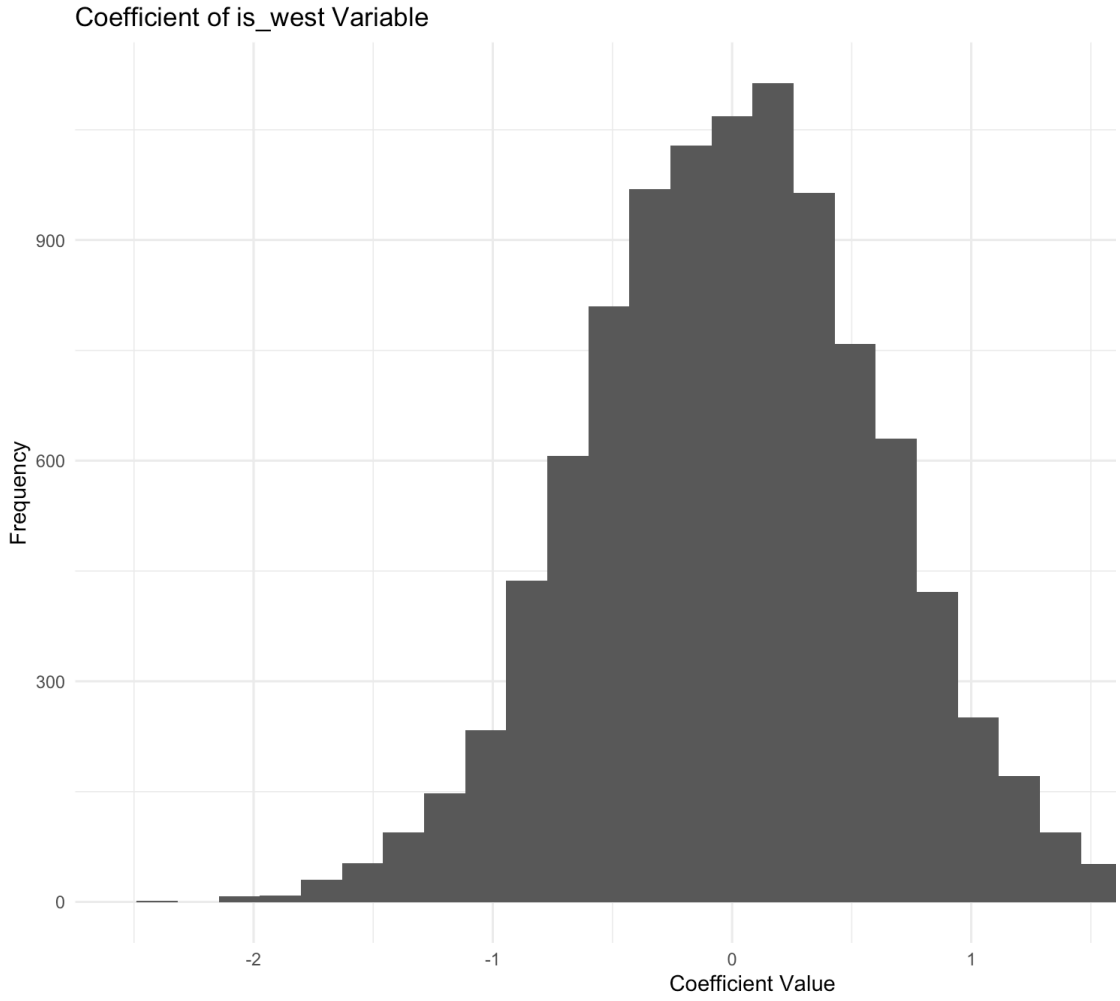
```
standings$random_conference <- standings$conference
coef_random_is_west <- c()
for (iter in 1:10000)
{
  # random order the conference label, 15 teams as West and 15 as East
  # sample the conference variable in the first 30 rows
  standings$random_conference[1:30] <- sample(standings$conference[1:30])

  # based on the first 30 rows (season 2020) to decide each team in previous seasons shal
  for (row_ind in 31:480)
  {
    standings$random_conference[row_ind] <- standings$random_conference[standings$season
_short == standings$team_short[row_ind]]
  }
  # regression model
  standings$random_is_west <- FALSE
  standings$random_is_west[standings$random_conference == "West"] <- TRUE
  sum(standings$random_is_west)

  model4 <- glm(playoffs_num ~ win_pct + random_is_west, data = standings, family = binom
  summary(model4)
  # extract the cofficient of is_west
  coef_random_is_west <- append(coef_random_is_west, model4$coefficients[3])
}
df_random_is_west <- data.frame(coef_random_is_west)
```

PART (a): Make a ggplot comparing these 10,000 randomized `is_west` coefficients to the `is_west` coefficient

ANSWER 14 (a):

```
ggplot(df_random_is_west, aes(x=coef_random_is_west)) +
  geom_histogram() +
  labs(title="Coefficient of is_west Variable",
            x="Coefficient Value",
            y="Frequency") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



PART (b): What do you conclude from your plot? Please answer in no more than 3 sentences.

ANSWER 14 (b): The value of the coefficient of the is_west variable appears to follow a normal distribution with a that in the conditions of randomly reordering the conference label and running numbers of iterations, whether a te not nearly has no impact on whether the team can make the playoffs.

Short Answer (Modeling)

Question 15

Two rookies come into the league and each play the same number of minutes. Rookie A shot 37/100 from 3 and r general manager asks you which player you expect to be a better three point shooter long term. You have a week project.

PART (a): What kind of data and information would you collect in order to answer this? Describe the features you build to answer the question. You don't need to actually do the work here, just describe your process. Please limi 5 sentences.

ANSWER 15 (a): The data I would collect include 1) the shooting location for the made and missed three-pointers player relies more on the corner threes or equally distributes his shots, 2) how many of the shots made were in a t difference with the opponent is within five or less points, 3) the assisted three-pointer field goals percentage, whic create his own shot, and 4) if the made shot is done in wide-open or being contested. I would consider applying a model with at least all of the above four factors included as predictors. The response variable for my model could a season or the average point the team could score in a game. That is because I always believe that on evaluating shooter or not, we shall always consider how can he help the team win.

PART (b): If you had to choose today without doing any further research, would you pick player A or player B? Wh to no more than 2 sentences.

ANSWER 15 (b): Player A will be my choice, as the data appears to tell me he is more capable to create his own much difference in the field goal percentage. However, if I could, I would strongly suggest not to evaluate a player data.

Question 16

QUESTION: You are studying offensive rebounding as part of your job as Data Analyst with an NBA team. You are will help predict when there will be an offensive rebound.

Your first model is a GBM (gradient boosted machine) that includes the location data (at the time the ball hits the on the floor. It outputs a prediction for the percentage chance each player will collect the rebound.

Your second model is a logistic regression model that includes the shot distance, the number of players who cras to the rim for the closest two offensive and defensive players. This model outputs a prediction for the percentage secures the rebound.

In no more than 4 sentences, how would you decide which model to move forward with? Why?

ANSWER 16: I might go with GBM as the response variable it delivers, the percentage of each player will collect t flexible for coaches to plan the team's second chance offense or transition defense. Coaches can base on each p rebound and his shooting range/defensive ability to best decide who shall go for the offensive rebound and who s However, I would prefer trying more predictors, e.g., the player's speed or how high/quick a player can jump, for b final call if I am really given the responsibility.

Question 17

QUESTION: Your coach hears about the project and is interested in the tradeoffs between offensive rebounding a you to expand the research you have been doing into a study to help him determine when/how many players he s offensive boards.

How would you use one of the models described above to answer the question for the coach? (Please select one about for this portion of the question.) What other research would be necessary to effectively answer this questio general terms, your plan to answer the coach's question and what you would plan to present to him. Please answ 6 sentences.

ANSWER 17: In addition to the GBM model mentioned above, I would also build two models for predicting the sc following second chance offense and the scores we might allow in the following transition defense. These two mo models with each of the 29 teams in the league as the group variable. I would present to the coach with the net s this position, which is the points gained in the second chance offense minus the points allow in the potential trans the coach to go with the strategy that results in the highest net scores for the team. The details of my strategy sha players, e.g., based on their positions, shall go for the offensive rebound, who shall stand by at the three-point lin transition defense.