Problem Set 6

Multivariate Analysis and Uncertainty

[YOUR NAME]

Due Date: 2024-02-23

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...` . Accept defaults and save this file as `[LAST NAME]_ps6.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps6.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `game_summary.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/game_summary.Rds).

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 8 total points, plus two extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compiled the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Also note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _____________________

Question 0

Require `tidyverse` and load the `game_summary.rds` (https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/4_Uni_Multivariate/data/game_summary.Rds?raw=true') data to an object called `games` .

```r
require(tidyverse)
```

```
## Loading required package: tidyverse
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
games <- read_rds('https://github.com/jbisbee1/DS1000_S2024/blob/main/data/game_summary.Rds?raw=
true')
```

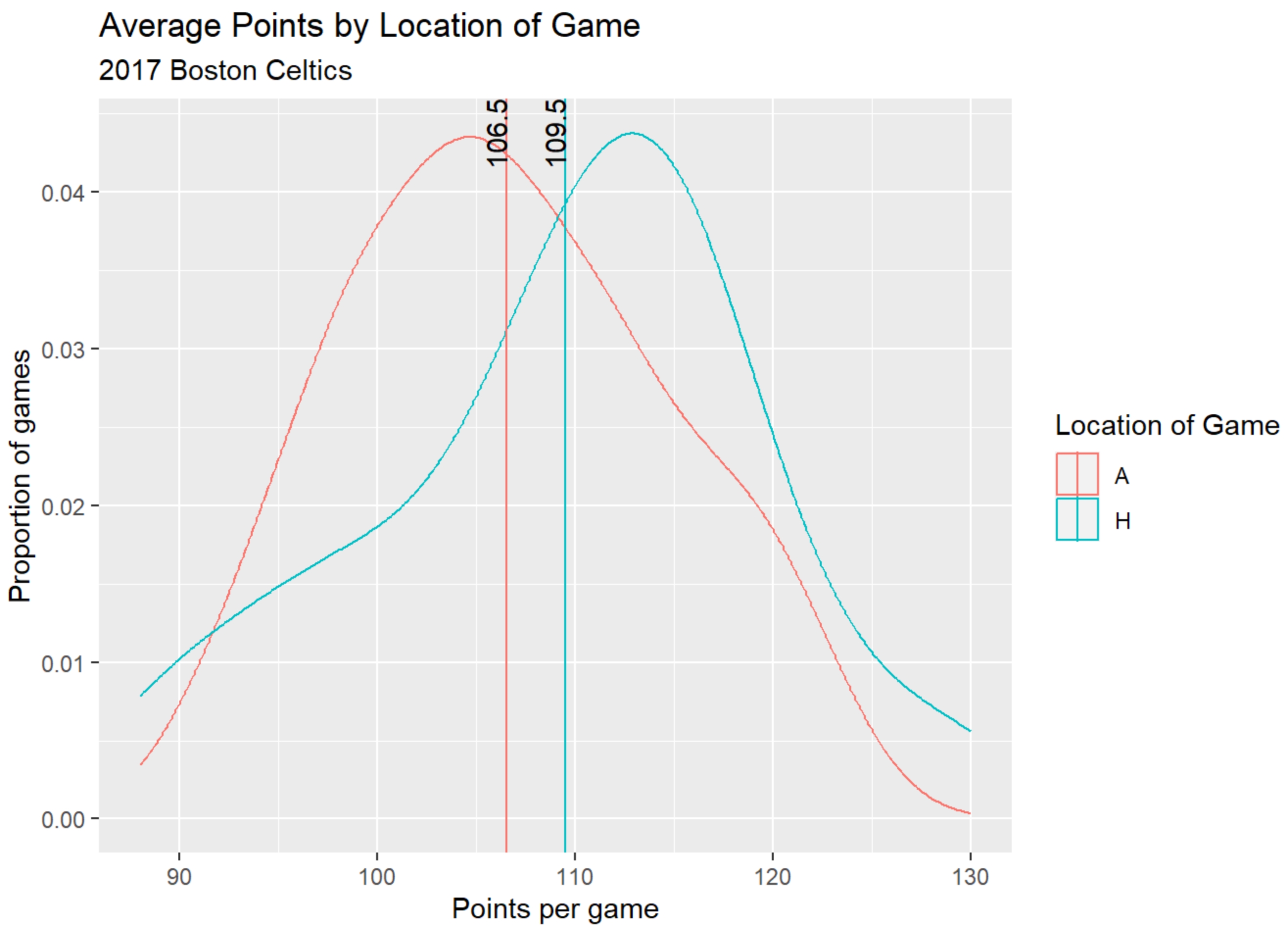
Question 1 [2 points]

How many points, on average, did the Boston Celtics score at home and away games in the 2017 season? Calculate this answer and also plot the multivariate relationship. Explain why your chosen visualization is justified. Draw two vertical lines for the average points at home and away.

```
(vertLines <- games %>%
filter(yearSeason == 2017,
       nameTeam == 'Boston Celtics') %>% # Filter to the 2017 season (yearSeason) AND to the Bos
ton Celtics (nameTeam)
  group_by(locationGame) %>% # Group by the location of the game (locatoinGame)
  summarise(avg_pts = mean(pts,na.rm=T))) # Calculate the average points (pts)
```

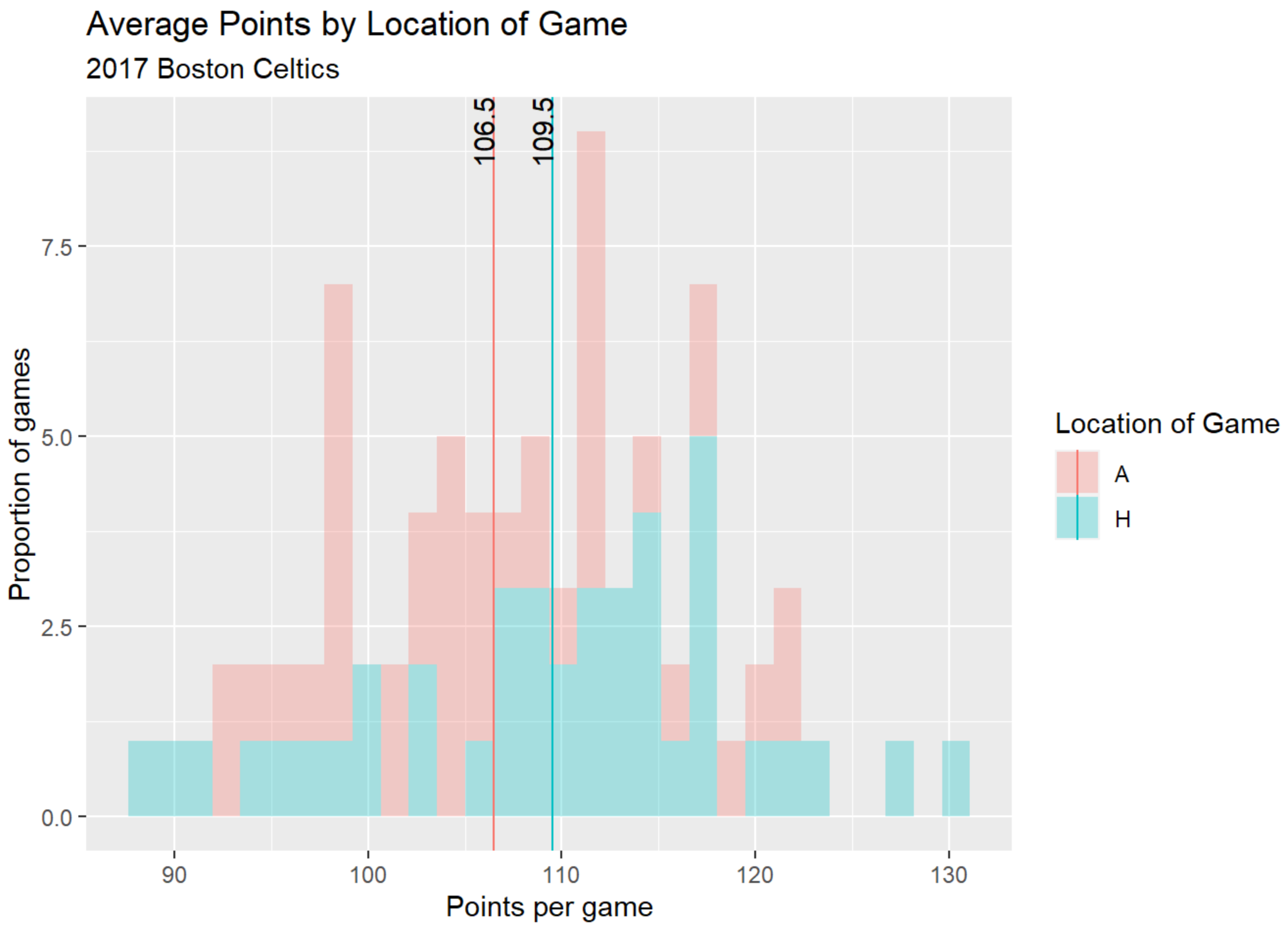
```
## # A tibble: 2 × 2
##   locationGame avg_pts
##   <chr>          <dbl>
## 1 A               107.
## 2 H               110.
```

```
games %>%
  filter(yearSeason == 2017,
         nameTeam == 'Boston Celtics') %>% # Filter to the 2017 season (yearSeason) AND to the Boston Celtics (nameTeam)
  ggplot(aes(x = pts,color = locationGame)) + # Create a multivariate plot comparing points scored between home and away games
  geom_density() + # Choose the appropriate geom_... for this plot (i.e., geom_histogram(), geom_density(), geom_bar(), etc.)
  labs(title = 'Average Points by Location of Game', # Add clear descriptions for the title, subtitle, axes, and legend
       subtitle = '2017 Boston Celtics',
       x = 'Points per game',
       y = 'Proportion of games',
       color = 'Location of Game') +
  geom_vline(data = vertLines,aes(xintercept = avg_pts,color = locationGame)) + # EC: add vertical lines for the average points scored at home and away.
  annotate(geom = 'text',x = vertLines$avg_pts,y = Inf,label = round(vertLines$avg_pts,1),vjust = 0,hjust = 1,angle = 90) # EC: label the vertical lines
```



```
games %>%
  filter(yearSeason == 2017,
         nameTeam == 'Boston Celtics') %>%
  ggplot(aes(x = pts,fill = locationGame)) +
  geom_histogram(alpha = .3) +
  labs(title = 'Average Points by Location of Game',
       subtitle = '2017 Boston Celtics',
       x = 'Points per game',
       y = 'Proportion of games',
       fill = 'Location of Game',
       color = 'Location of Game') +
  geom_vline(data = vertLines,aes(xintercept = avg_pts,color = locationGame)) +
  annotate(geom = 'text',x = vertLines$avg_pts,y = Inf,label = round(vertLines$avg_pts,1),vjust
= 0,hjust = 1,angle = 90)
```

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



I chose a geom_density that was colored by the location of the game. I could have also chosen a histogram.

Question 2 [2 points]

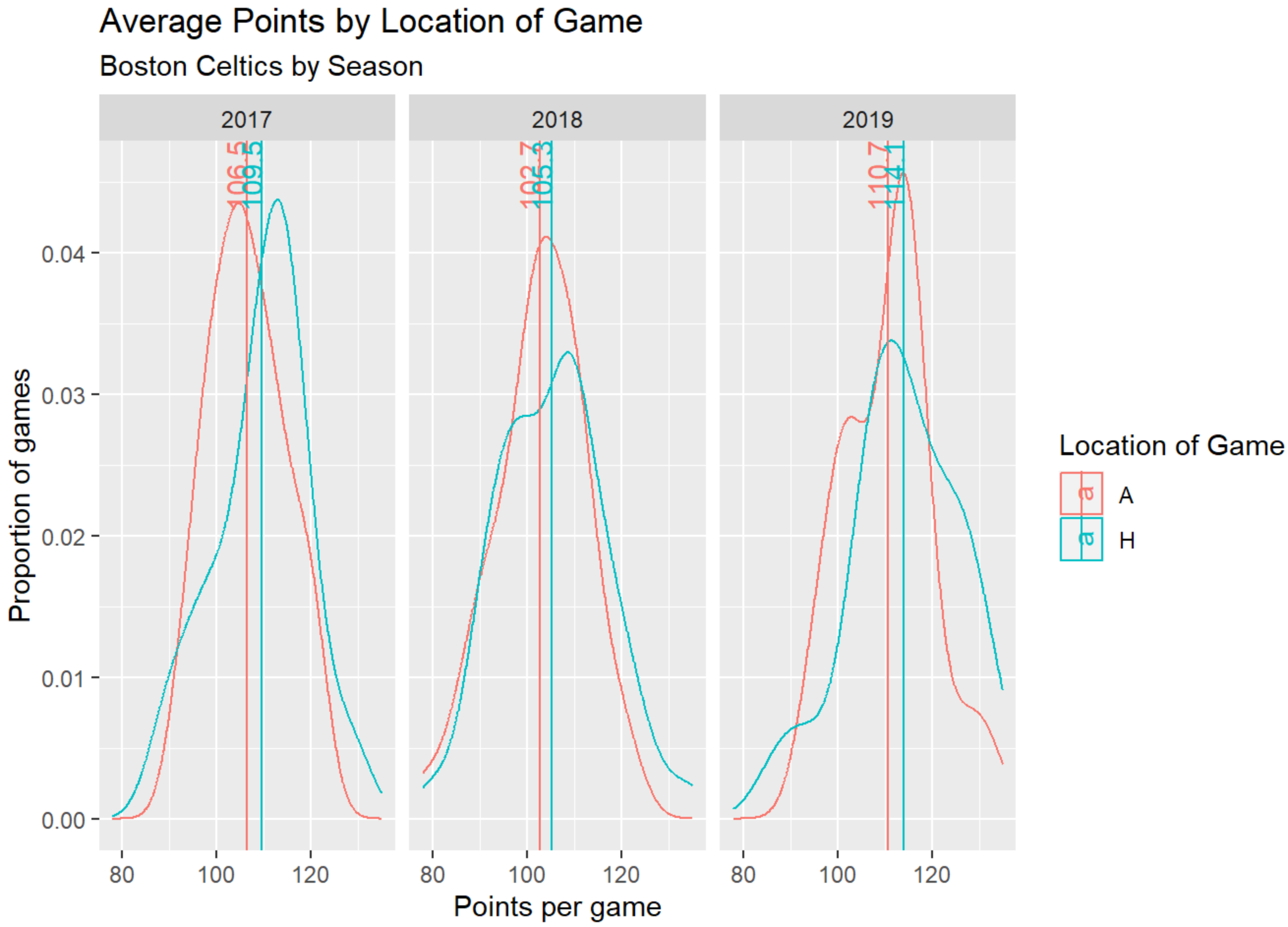
Now recreate the same plot for the 2018, 2019, and combined seasons. Imagine that you work for the Celtics organization and Brad Stevens (the GM), asks you if the team scores more points at home or away? Based on your analysis, what would you tell him?

```
# By season
(vertLines <- games %>%
filter(nameTeam == 'Boston Celtics') %>% # Filter to the Boston Celtics (nameTeam)
  group_by(locationGame,yearSeason) %>% # Group by the location (locationGame) and the season (y
earSeason)
  summarise(avg_pts = mean(pts,na.rm=T))) # Calculate the average points (pts)
```

```
## `summarise()` has grouped output by 'locationGame'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 6 × 3
## # Groups:   locationGame [2]
##   locationGame yearSeason avg_pts
##   <chr>             <int>   <dbl>
## 1 A                  2017    107.
## 2 A                  2018    103.
## 3 A                  2019    111.
## 4 H                  2017    110.
## 5 H                  2018    105.
## 6 H                  2019    114.
```

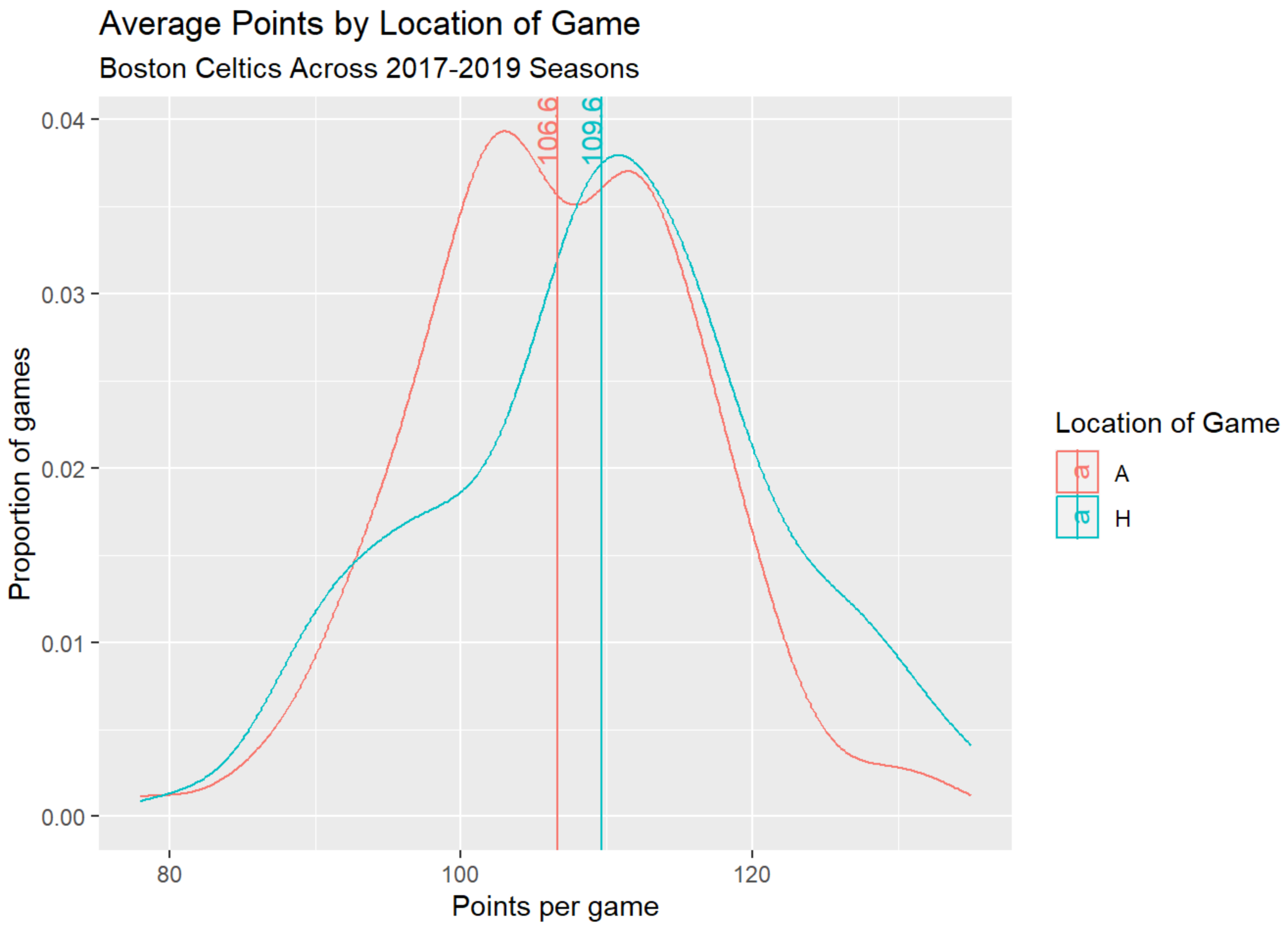
```
games %>%
  filter(nameTeam == 'Boston Celtics') %>% # Filter to the 2017 season (yearSeason) AND to the B
oston Celtics (nameTeam)
  ggplot(aes(x = pts,color = locationGame)) + # Create a multivariate plot comparing points scor
ed between home and away games
  geom_density() + # Choose the appropriate geom_... for this plot (i.e., geom_histogram(), geom
_density(), geom_bar(), etc.)
  labs(title = 'Average Points by Location of Game', # Add clear descriptions for the title, sub
title, axes, and legend
       subtitle = 'Boston Celtics by Season',
       x = 'Points per game',
       y = 'Proportion of games',
       color = 'Location of Game') +
  facet_wrap(~yearSeason) + # Create separate panels for each season (facet_wrap())
  geom_vline(data = vertLines,aes(xintercept = avg_pts,color = locationGame)) +
  geom_text(data = vertLines,aes(x = avg_pts,y = Inf,color = locationGame,label = round(avg_pts,
1)),
            vjust = 0,hjust = 1,angle = 90)
```



```
# Over all seasons combined
(vertLines <- games %>%
filter(nameTeam == 'Boston Celtics') %>% # Filter to the Boston Celtics (nameTeam)
  group_by(locationGame) %>% # Group by the location (locationGame)
  summarise(avg_pts = mean(pts,na.rm=T))) # Calculate the average points (pts)
```

```
## # A tibble: 2 × 2
##   locationGame avg_pts
##   <chr>          <dbl>
## 1 A               107.
## 2 H               110.
```

```
games %>%
  filter(nameTeam == 'Boston Celtics') %>% # Filter to the 2017 season (yearSeason) AND to the B
oston Celtics (nameTeam)
  ggplot(aes(x = pts,color = locationGame)) + # Create a multivariate plot comparing points scor
ed between home and away games
  geom_density() + # Choose the appropriate geom_... for this plot (i.e., geom_histogram(), geom
_density(), geom_bar(), etc.)
  labs(title = 'Average Points by Location of Game', # Add clear descriptions for the title, sub
title, axes, and legend
       subtitle = 'Boston Celtics Across 2017-2019 Seasons',
       x = 'Points per game',
       y = 'Proportion of games',
       color = 'Location of Game') +
  geom_vline(data = vertLines,aes(xintercept = avg_pts,color = locationGame)) +
  geom_text(data = vertLines,aes(x = avg_pts,y = Inf,color = locationGame,label = round(avg_pts,
1)),
            vjust = 0,hjust = 1,angle = 90)
```



> The Celtics scored more points at home games than away games for every season in the data, as well as when combining all the seasons together. Based on this analysis, I would tell Brad Stevens that the Celtics score more points at home games than at away games. Overall, the difference is equivalent to roughly one 3-point shot: 106.6 points at away games and 109.6 points at home games.

Question 3 [2 points]

Brad Stevens thanks you for your answer, but is a well-trained statistician in his own right, and wants to know how confident you are in your claim. Bootstrap sample the data 1,000 times to provide him with a more sophisticated answer. How confident are you in your conclusion that the Celtics score more points at home games than away games? Make sure to `set.seed(123)` to ensure you get the same answer every time you `knit` your code!

```
set.seed(123) # Set the seed!
forBS <- games %>% # To make things easier, create a new data object that is filtered to just the Celtics
    filter(nameTeam == 'Boston Celtics') # Filter to the Celtics (nameTeam)

bsRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:1000) { # Loop 1,000 times
  bsRes <- forBS %>%
    sample_n(size = nrow(forBS),replace = T) %>% # Sample the data with replacement using all possible rows
    group_by(locationGame) %>% # Group by the location of the game (locationGame)
    summarise(avg_pts = mean(pts,na.rm=T)) %>% # Calculate the average points (pts)
    ungroup() %>% # Best practices!
    spread(locationGame,avg_pts) %>% # Spread the data to get one column for average points at home and another for average points away
    mutate(diff = H - A, # Calculate the difference between home and away points
           bsInd = i) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 133
}

# Calculate the confidence
bsRes %>%
  summarise(confidence = mean(diff > 0), # Calculate the proportion of bootstrap simulations where the home points are greater than the away points
            avg_diff = mean(diff)) # Calculate the overall average difference
```

```
## # A tibble: 1 × 2
##   confidence avg_diff
##        <dbl>    <dbl>
## 1      0.992     2.93
```

> I am 99.2% confident in my conclusion that the Celtics score more points at home games than away games. Furthermore, the average difference is just about 3 points (2.93) over the 1,000 bootstrapped simulatoins.

Question 4 [2 points]

Re-do this analysis for three other statistics of interest to Brad: total rebounds (treb), turnovers (tov), and field goal percent (pctFG). Do you notice anything strange in these results? What might explain it?

```r
bsRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:1000) { # Loop 1,000 times
  bsRes <- forBS %>%
    sample_n(size = nrow(forBS),replace = T) %>% # Sample the data with replacement using all possible rows
    group_by(locationGame) %>% # Group by the location of the game (locationGame)
    summarise(avg_reb = mean(treb,na.rm=T), # Calculate the average total rebounds (treb)
              avg_tov = mean(tov,na.rm=T), # Calculate the average turnovers (tov)
              avg_pctFG = mean(pctFG,na.rm=T)) %>% # Calculate the average field goal shooting percentage (pctFG)
    ungroup() %>% # Best practices!
    pivot_wider(names_from = locationGame, # Pivot wider to get each measure in its own colunm for homme and away games
                values_from = c('avg_reb','avg_tov','avg_pctFG')) %>% # Use the values from the variables you created above
    mutate(diff_reb = avg_reb_H - avg_reb_A, # Calculate the difference between home and away total rebounds
           diff_tov = avg_tov_H - avg_tov_A, # Calculate the difference between home and away turnovers
           diff_pctFG = avg_pctFG_H - avg_pctFG_A, # Calculate the difference between home and away field goal percentages
           bsInd = i) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 165
}

# Calculate the confidence
bsRes %>%
  summarise(confidence_reb = mean(diff_reb > 0),
            confidence_tov = mean(diff_tov > 0),
            confidence_pctFG = mean(diff_pctFG > 0))
```

```
## # A tibble: 1 × 3
##   confidence_reb confidence_tov confidence_pctFG
##            <dbl>          <dbl>            <dbl>
## 1          0.994          0.923            0.885
```

> I am 99.4% confident that the Celtics rebound more at home games than away games. I am 92.3% confident that they turn over the ball more at home games than away games. And I am 88.5% confident that they shoot more accurately at home than away games. These results are surprising since turnovers are theoretically bad for a basketball team, yet we find that the Celtics have more turnovers at home games than away games. This might be due to a faster pace of play, where the Celtics move the ball around more, providing more opportunityies for points and rebounds, but also more turnovers.

Extra Credit [2 points]

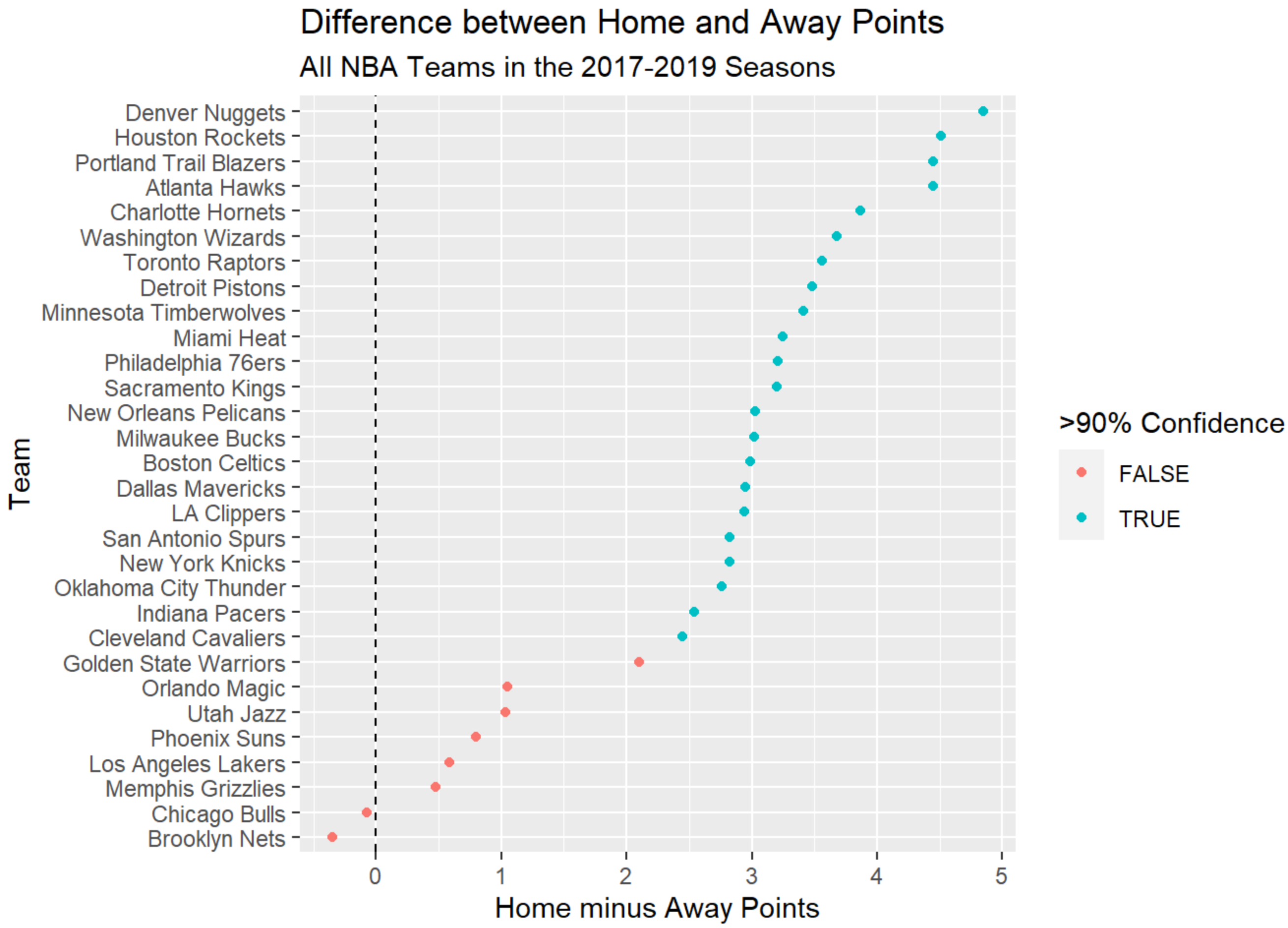
Now Brad is asking for a similar analysis of other teams. Calculate the difference between home and away points for every team in the league and prepare a summary table that includes both the average difference for each team, as well as your confidence about whether the difference is not zero. Based on these data, would you argue that there is an **overall** home court advantage in terms of points across the NBA writ large? Visualize these summary results by plotting the difference on the x-axis, the teams (reordered) on the y-axis, and the points colored by whether you are more than 90% confident in your answer. How should we interpret confidence levels less than 50%?

```
bsRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:1000) { # Loop 1,000 times
  bsRes <- games %>%
    group_by(nameTeam) %>%
    sample_n(size = n(),replace = T) %>% # Sample the data with replacement using all possible rows
    group_by(locationGame,nameTeam) %>% # Group by the location of the game (locationGame)
    summarise(avg_pts = mean(pts,na.rm=T),.groups = 'drop') %>% # Calculate the average turnovers (tov)
    pivot_wider(id_cols = nameTeam,
                names_from = locationGame, # Pivot wider to get each measure in its own column for homme and away games
                values_from = c('avg_pts')) %>% # Use the values from the variables you created above
    mutate(diff = H - A, # Calculate the difference between home and away turnovers
           bsInd = i) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 165
}

(toplot <- bsRes %>%
  group_by(nameTeam) %>%
  summarise(conf = round(mean(diff > 0),2),
            diff = round(mean(diff),2)))
```

```
## # A tibble: 30 × 3
##    nameTeam               conf  diff
##    <chr>                 <dbl> <dbl>
##  1 Atlanta Hawks          1     4.45
##  2 Boston Celtics         0.99  2.99
##  3 Brooklyn Nets          0.43 -0.35
##  4 Charlotte Hornets      0.99  3.87
##  5 Chicago Bulls          0.5  -0.07
##  6 Cleveland Cavaliers    0.92  2.45
##  7 Dallas Mavericks       0.98  2.95
##  8 Denver Nuggets         1     4.85
##  9 Detroit Pistons        0.99  3.48
## 10 Golden State Warriors  0.88  2.1
## # ℹ 20 more rows
```

```
toplot %>%
  ggplot(aes(x = diff,y = reorder(nameTeam,diff),color = conf > .9 | conf < .1))+
  geom_point() +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  labs(title = 'Difference between Home and Away Points',
       subtitle = 'All NBA Teams in the 2017-2019 Seasons',
       x= 'Home minus Away Points',
       y = 'Team',
       color = '>90% Confidence')
```



> Here we find much stronger evidence that teams generally score more points at home than away games across the NBA. Every team except the Bulls and Nets score more points at home than away, and the majority of these differences we can confidently say are greater than zero at the 90% level. Confidence levels less than 50% mean that teams scored more home points than away points in fewer than 50% of simulated realities. This is equivalent to saying that they scored more away points than home points in more than 50% of simulated realities. In other words, if the confidence is less than 0.5, we can flip the statement and say we are 1-the confidence.