Midterm Exam

Prof. Bisbee

Due Date: 2022/10/16 @ 11:59PM CST

Getting Set Up

Download the zipped `midterm` folder from Github and save to your class folder for DS1000. Unzip it, retaining the folder structure. Within you should fine a `code` folder and a `data` folder. The `data` folder should contain the two datasets we will be using in the midterm: `sc_debt.Rds` and `game_summary.Rds`. Within the `code` folder is this very file: `DS1000_midterm_exam.Rmd`.

Rename this file to `[LAST NAME]_midterm.Rmd` and do all your work in this file.

All of the following questions should be answered with `set.seed(123)`. To start, require `tidyverse` and load the `sc_debt.Rds` data to `debt`.

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
debt <- readRDS('../data/sc_debt.Rds')
```

NB: you will be penalized 1 point for each poorly labeled figure. You will be penalized 5 points if the submission is not a PDF of the knitted output. The total points available are 50.

Question 0: Independent Work Statement

Please sign your name in the space provided by typing out your full name in place of the underline:

"I, _____, am aware of the serious nature of plagiarism and affirm that I did not collaborate with other students while completing this midterm."

Question 1 [5 points]

Describe the following three variables: `grad_debt_mdn`, `costt4_a`, and `region`. What `class` are they? Do they contain any missing data? How should you visualize each of them for univariate description? Plot your choices with clear labels.

```
################ grad_debt_mdn ######################
# Looking first
debt %>% select(grad_debt_mdn)
```

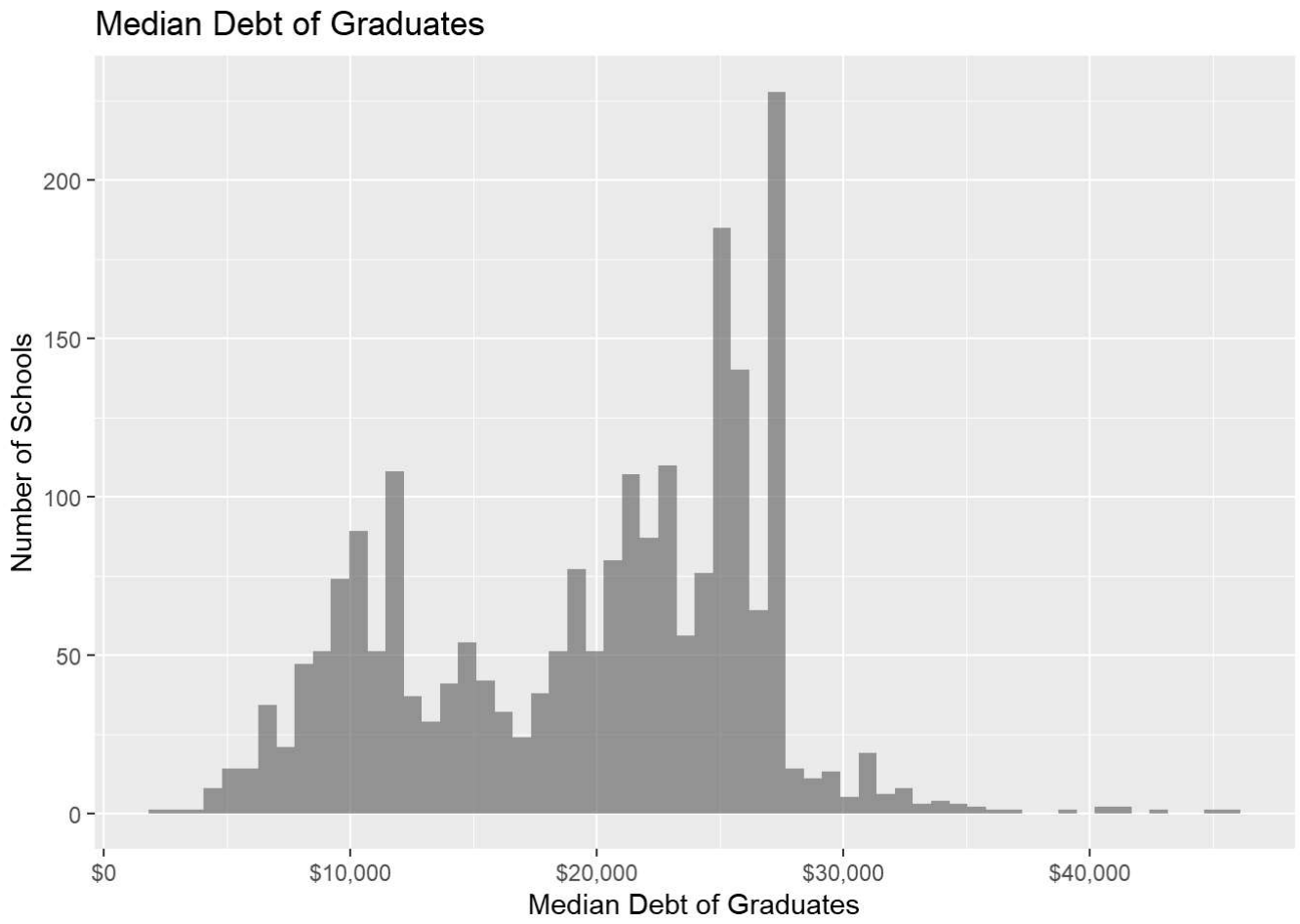
```
## # A tibble: 2,546 × 1
##    grad_debt_mdn
##            <int>
##  1         33375
##  2         22500
##  3         27334
##  4         21607
##  5         32000
##  6         23250
##  7         12500
##  8         19500
##  9         24826
## 10         21281
## # … with 2,536 more rows
```

```
# Summary second
debt %>%
select(grad_debt_mdn) %>%
summary()
```

```
##  grad_debt_mdn
##  Min.   : 2332
##  1st Qu.:13000
##  Median :21500
##  Mean   :19646
##  3rd Qu.:25125
##  Max.   :45881
##  NA's   :325
```

```
# Plotting third
debt %>%
  ggplot(aes(x = grad_debt_mdn)) +
  geom_histogram(alpha = .6,bins = 60) + # Or density
  labs(title = 'Median Debt of Graduates',
       x = 'Median Debt of Graduates',
       y = 'Number of Schools') +
  scale_x_continuous(labels = scales::dollar)
```

```
## Warning: Removed 325 rows containing non-finite values (stat_bin).
```




```
################# costt4_a ######################
# Looking first
debt %>% select(costt4_a)
```

```
## # A tibble: 2,546 × 1
##    costt4_a
##       <int>
##  1    23053
##  2    24495
##  3    14800
##  4    23917
##  5    21866
##  6    29872
##  7    10493
##  8       NA
##  9    19849
## 10    31590
## # … with 2,536 more rows
```

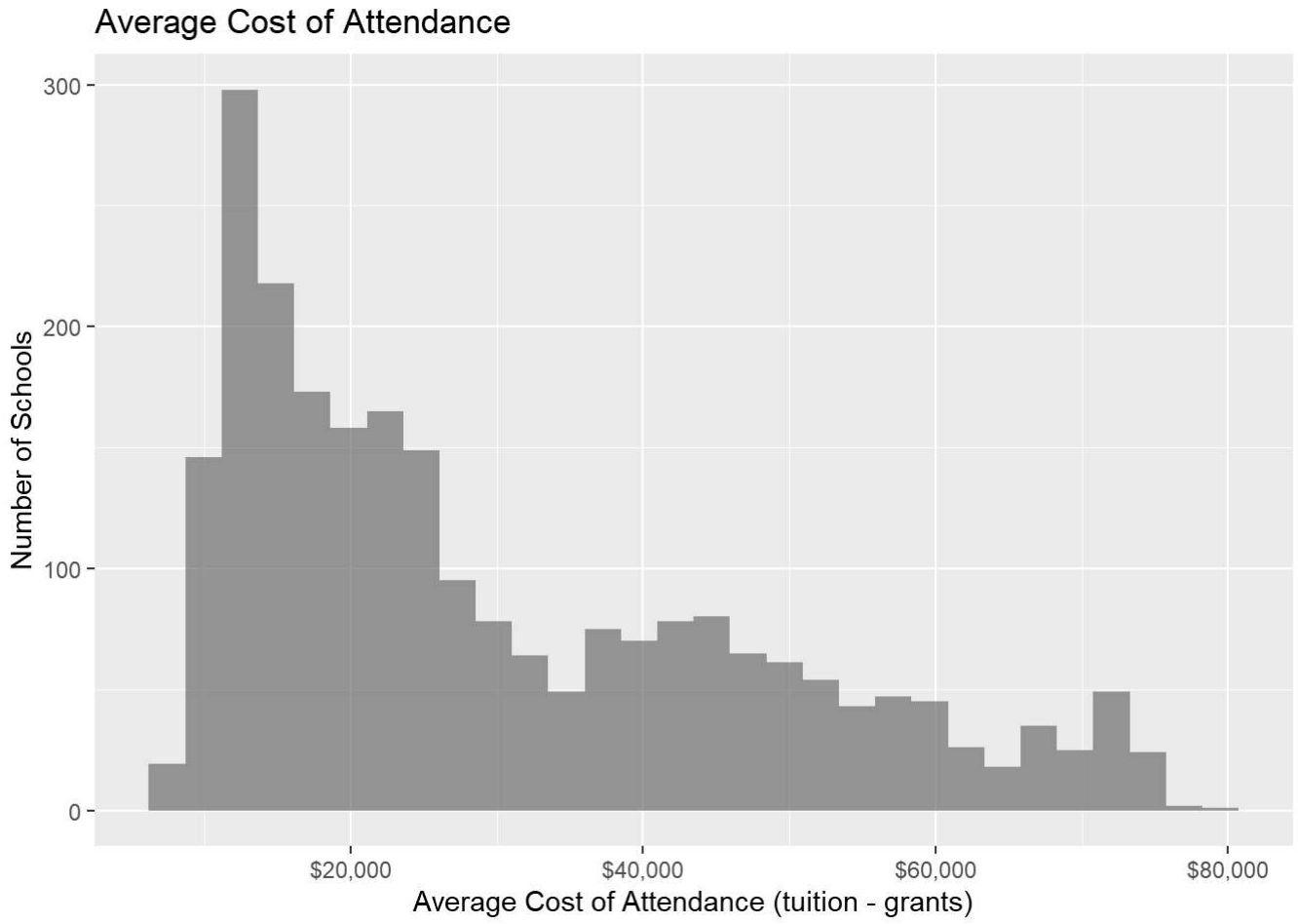
```
# Summary second
debt %>%
select(costt4_a) %>%
summary()
```

```
##     costt4_a
##  Min.   : 6525
##  1st Qu.:15051
##  Median :23948
##  Mean   :29971
##  3rd Qu.:42824
##  Max.   :78555
##  NA's   :136
```

```
# Plotting third
debt %>%
  ggplot(aes(x = costt4_a)) +
  geom_histogram(alpha = .6) + # Or density
  labs(title = 'Average Cost of Attendance',
       x = 'Average Cost of Attendance (tuition - grants)',
       y = 'Number of Schools') +
  scale_x_continuous(labels = scales::dollar)
```

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

```
## Warning: Removed 136 rows containing non-finite values (stat_bin).
```



```
################ region ######################
# Looking first
debt %>% select(region)
```

```
## # A tibble: 2,546 × 1
##    region
##    <chr>
##  1 Southeast
##  2 Southeast
##  3 Southeast
##  4 Southeast
##  5 Southeast
##  6 Southeast
##  7 Southeast
##  8 Southeast
##  9 Southeast
## 10 Southeast
## # … with 2,536 more rows
```

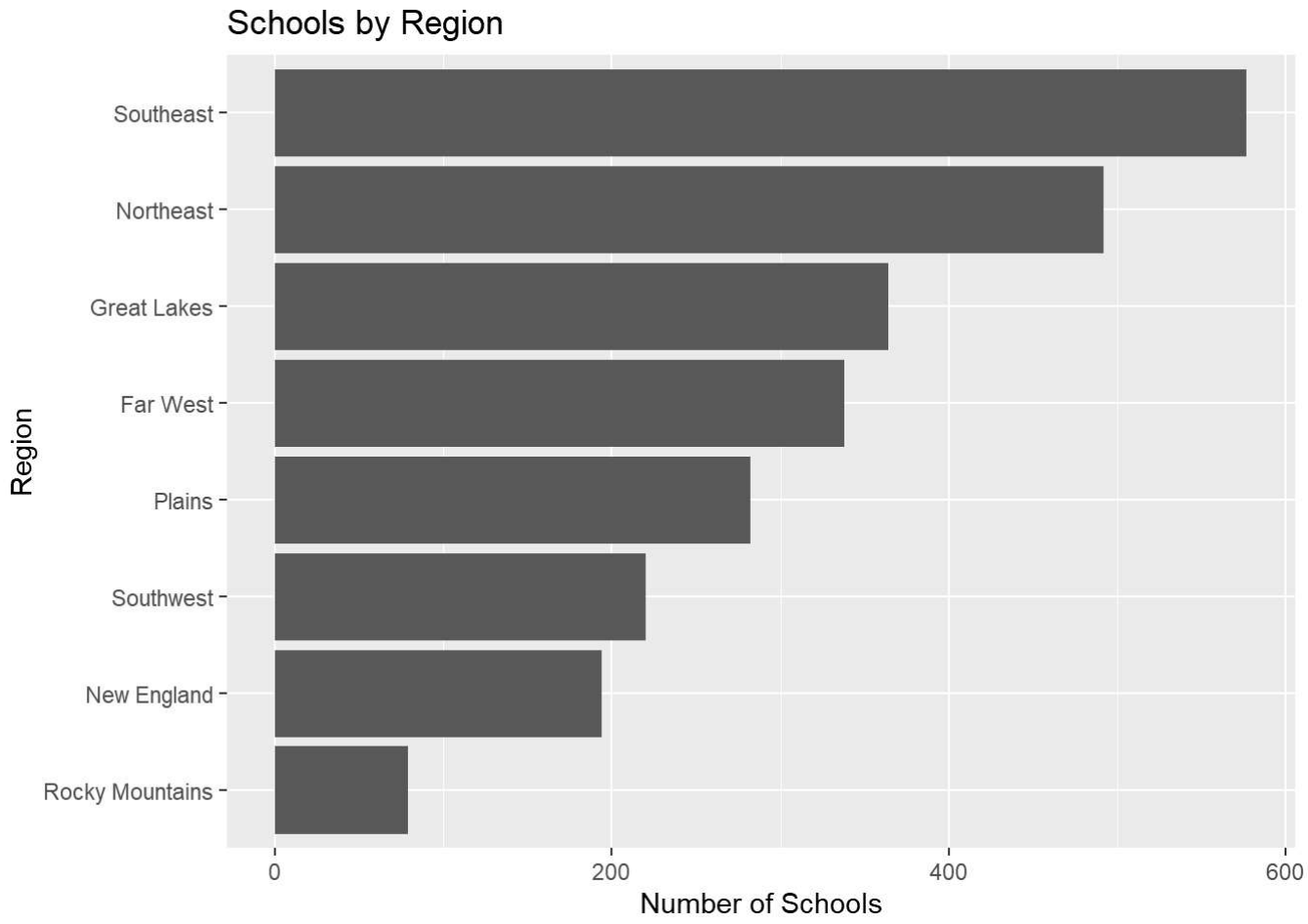
```
# Summary second
debt %>%
select(region) %>%
summary()
```

```
##     region
##  Length:2546
##  Class :character
##  Mode  :character
```

```
debt %>%
  count(region)
```

```
## # A tibble: 8 × 2
##   region              n
##   <chr>           <int>
## 1 Far West          338
## 2 Great Lakes       364
## 3 New England       194
## 4 Northeast         492
## 5 Plains            282
## 6 Rocky Mountains    79
## 7 Southeast         577
## 8 Southwest         220
```

```
# Plotting third
debt %>%
group_by(region) %>%
summarise(n=n()) %>%
  ggplot(aes(x = reorder(region,n),y = n)) +
  geom_bar(stat = 'identity') +
  labs(title = 'Schools by Region',
       x = 'Region',
       y = 'Number of Schools') +
       coord_flip()
```



- Median graduate debt is a continuous variable `dbl` with 325 missing values. We should visualize it with either a histogram or a density plot.

- Average cost of attendance is a continuous variable `dbl` with 136 missing values. We should visualize it with either a histogram or a density plot.

- Region is a categorical variable `chr` with no missing values. We should visualize it with a barplot.

Question 2 [5 points]

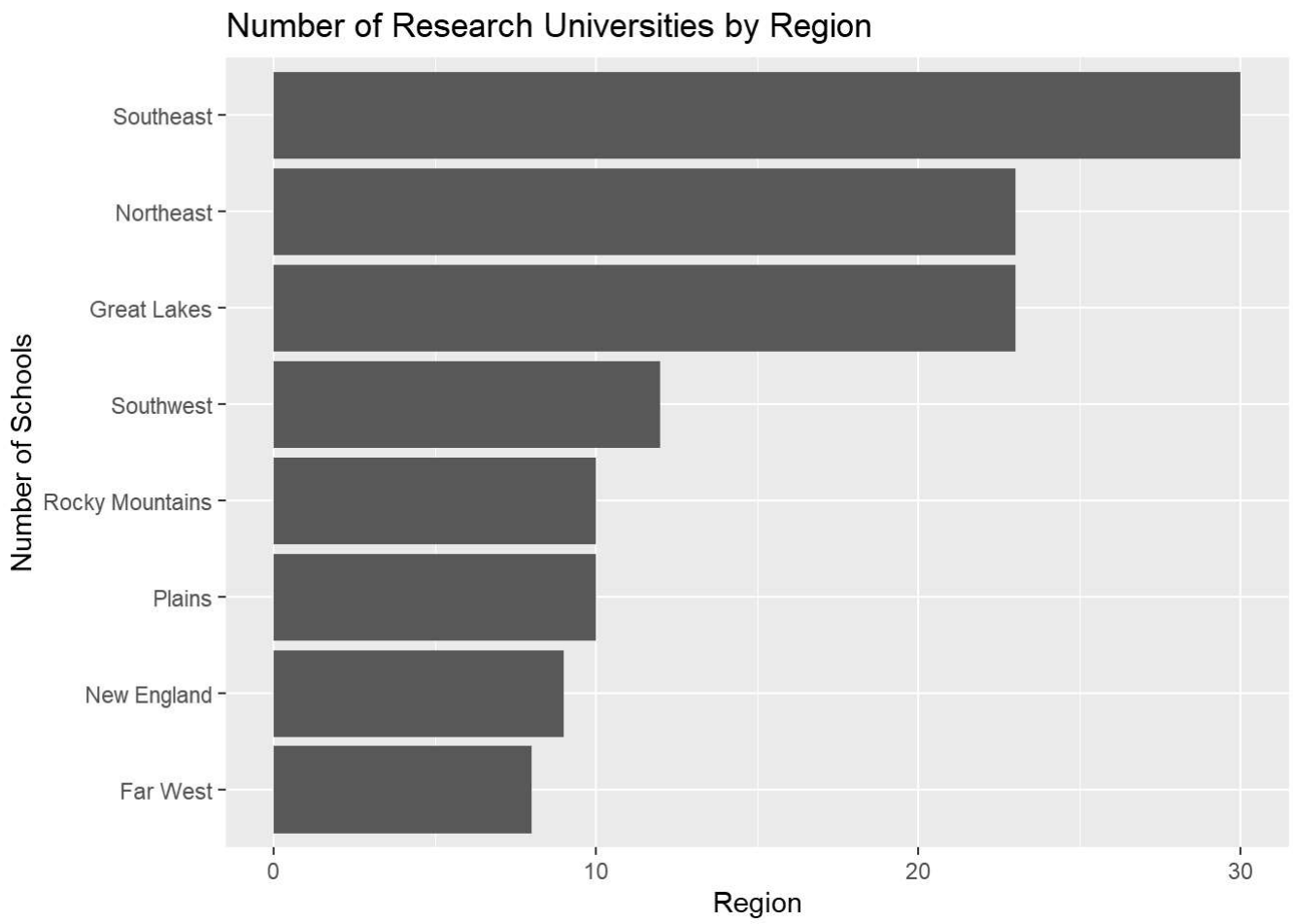
In which region are the most research universities located? Which region has the fewest? Plot your answer.

```
debt %>%
  group_by(region,research_u) %>%
  summarise(n=n()) %>%
  filter(research_u == 1) %>%
  arrange(-n)
```

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

```
## # A tibble: 8 × 3
## # Groups:   region [8]
##   region          research_u     n
##   <chr>                <dbl> <int>
## 1 Southeast                1    30
## 2 Great Lakes              1    23
## 3 Northeast                1    23
## 4 Southwest                1    12
## 5 Plains                   1    10
## 6 Rocky Mountains          1    10
## 7 New England              1     9
## 8 Far West                 1     8
```

```
debt %>%
  filter(research_u == 1) %>%
  count(region) %>%
  ggplot(aes(x = reorder(region,n),y = n)) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  labs(title = 'Number of Research Universities by Region',
       x = 'Number of Schools',
       y = 'Region')
```



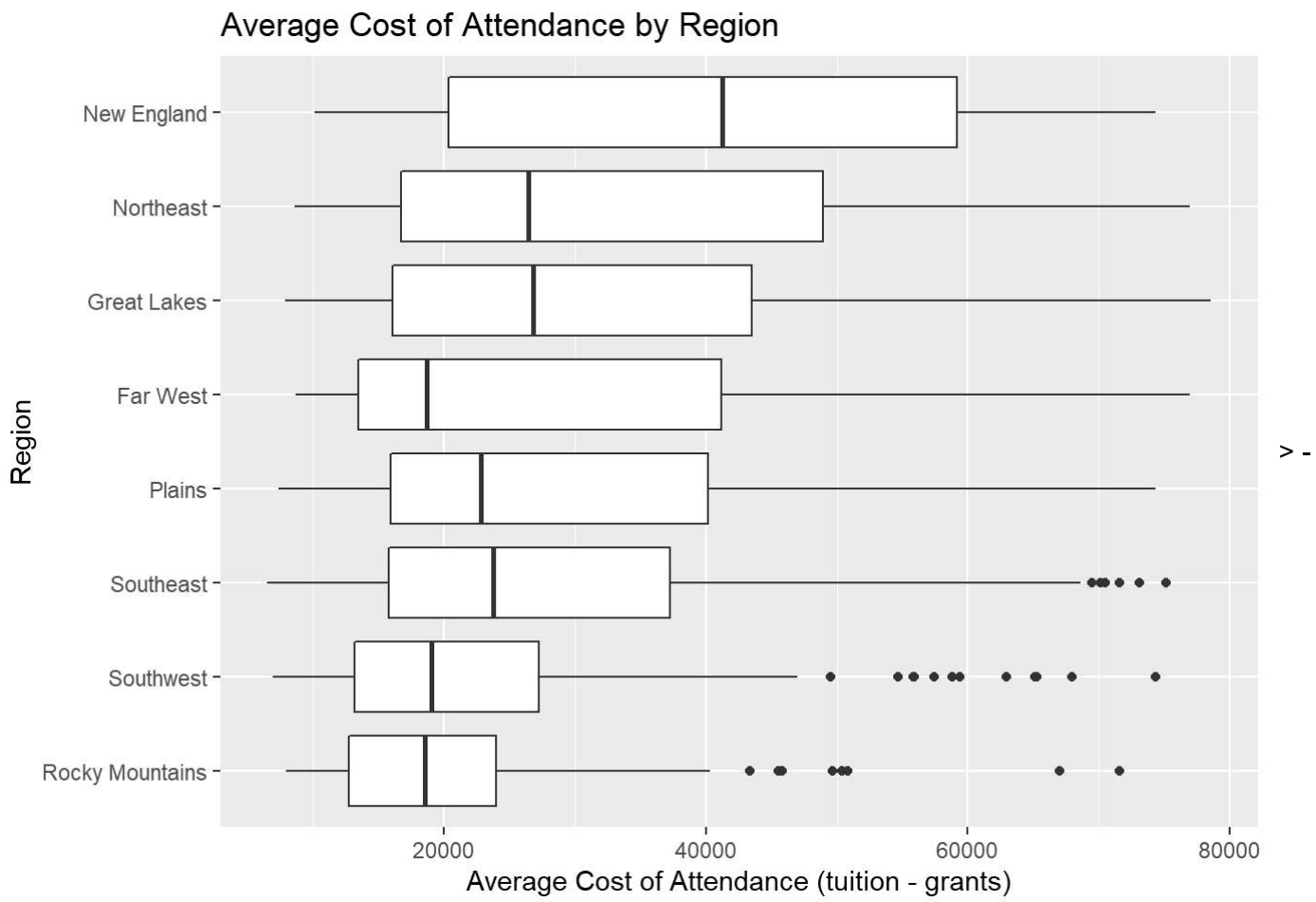
- The Southeast has the most research universities while the Far West has the fewest.

Question 3 [5 points]

What is the conditional relationship between region and average cost of attendance? Plot this relationship using the appropriate `geom_` function and justify your choice.

```
debt %>%
  ggplot(aes(x = reorder(region,costt4_a,na.rm = TRUE),y = costt4_a)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Average Cost of Attendance by Region",
       y = "Average Cost of Attendance (tuition - grants)",
       x = "Region")
```

```
## Warning: Removed 136 rows containing non-finite values (stat_boxplot).
```



New England has the most expensive schools whereas the Rocky Mountains have the least expensive schools. I choose to use a `geom_boxplot()` to visualize the conditional relationship because we are comparing a continuous measure (cost) to a categorical measure (region) which has more than 2 categories. I could have also chosen `geom_violin()`.

Question 4 [5 points]

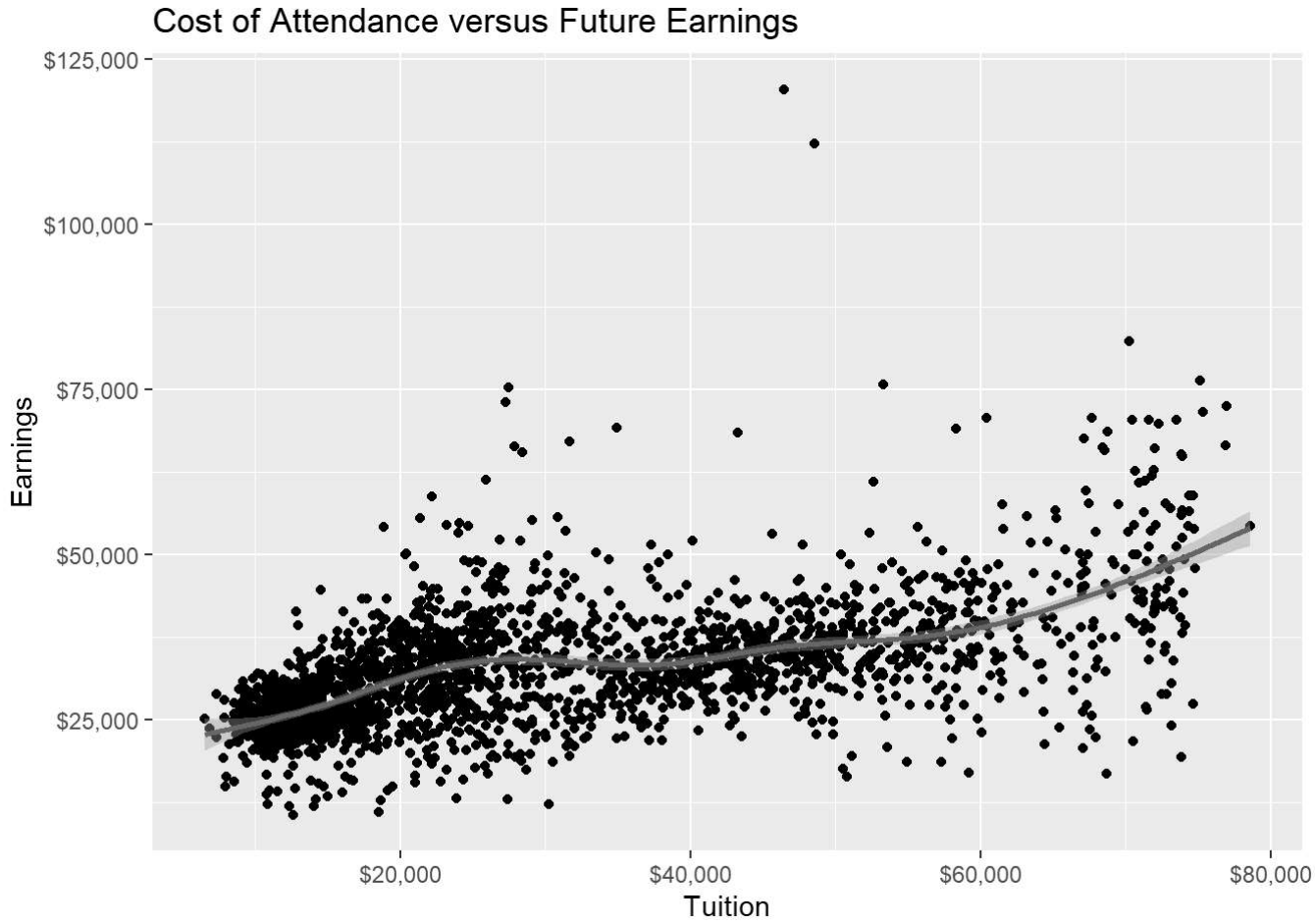
What is the conditional relationship between the cost of attendance and future earnings of graduates? As above, plot this using the best `geom_` and justify your choice. Highlight the relationship with `geom_smooth()`, and discuss what it means. Why might this correlation be spurious?

```
debt %>%
  ggplot(aes(x = costt4_a,y = md_earn_wne_p6)) +
  geom_point() +
  geom_smooth() +
  scale_x_continuous(label = scales::dollar) +
  scale_y_continuous(label = scales::dollar) +
  labs(title = "Cost of Attendance versus Future Earnings",
       x = "Tuition",
       y = "Earnings")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 307 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 307 rows containing missing values (geom_point).
```



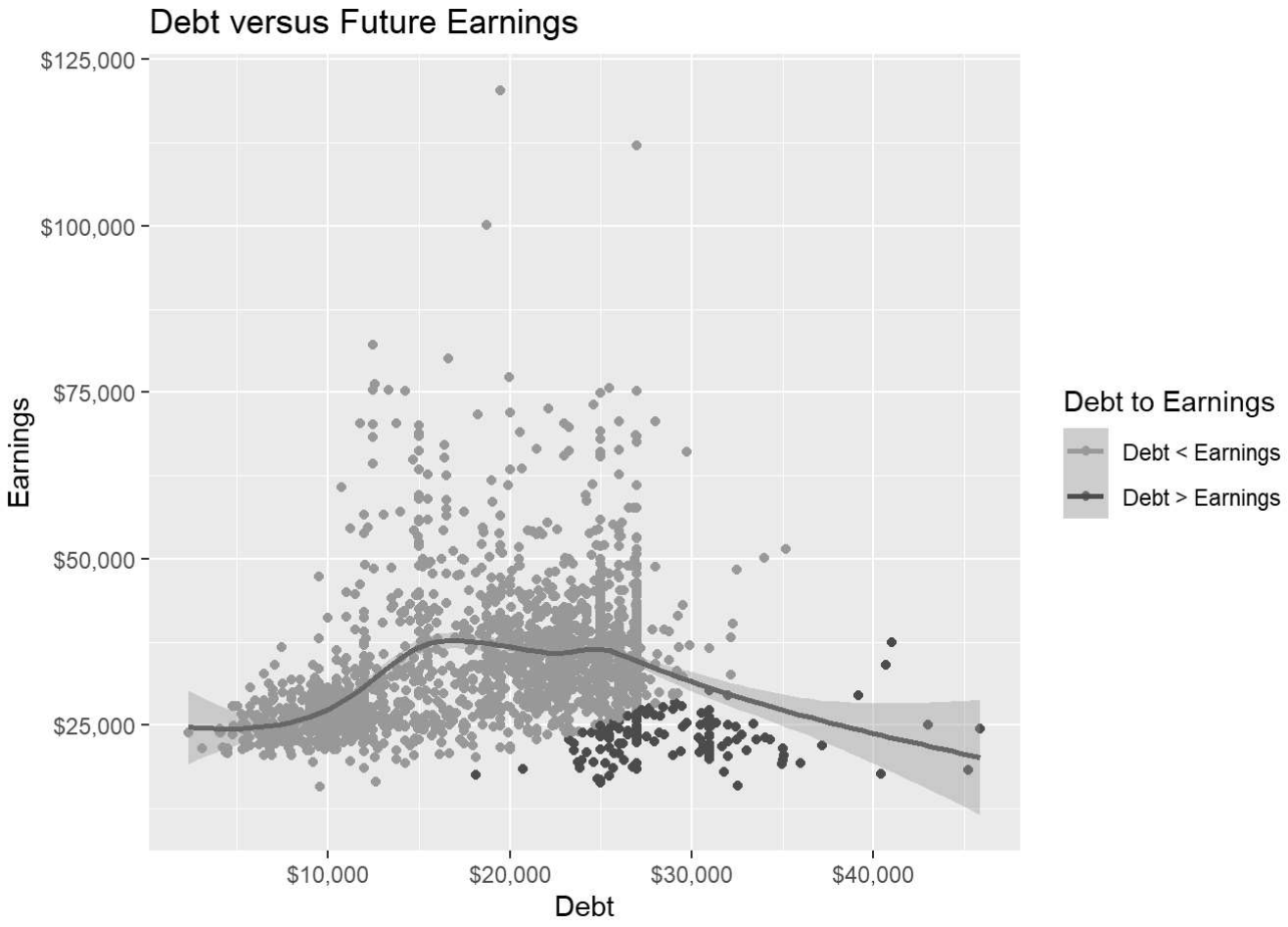
- There is a positive association between the cost of attendance and future earnings. This suggests that paying more for an expensive school leads to greater future earnings. However, this might be a spurious association if wealthier students can afford more expensive schools, and are also part of social networks that lead to higher paying jobs.

Question 5 [5 points]

What is the conditional relationship between the median debt of graduates and their future earnings? Identify those schools for whom the debt is larger than their earnings. Plot the conditional relationship using the appropriate `geom_` , coloring schools by whether the debt is greater than the earnings. To answer, you only need to generate the plot.

```
debt %>%
  mutate(debtGreater = grad_debt_mdn > md_earn_wne_p6) %>%
  filter(!is.na(debtGreater)) %>%
  ggplot(aes(x = grad_debt_mdn,y = md_earn_wne_p6,color = debtGreater,group = 1)) +
  geom_point() +
  geom_smooth() +
  scale_x_continuous(label = scales::dollar) +
  scale_y_continuous(label = scales::dollar) +
  labs(title = "Debt versus Future Earnings",
       x = "Debt",
       y = "Earnings") +
  scale_color_manual(name = 'Debt to Earnings',values = c('grey60','red'),
                     labels = c('Debt < Earnings','Debt > Earnings'))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Question 6 [5 points]

Now load the `game_summary.Rds` data to `gms`. Identify the team with the fewest points over all three seasons. Who is it and how many points did they score? Were they also the team with the lowest win percentage over all three seasons? If not, how far were they from the bottom of the pack? Now plot their points per game by season, using the best `geom_`. **NOTE:** You should convert `yearSeason` to a categorical variable with the `factor()` command.

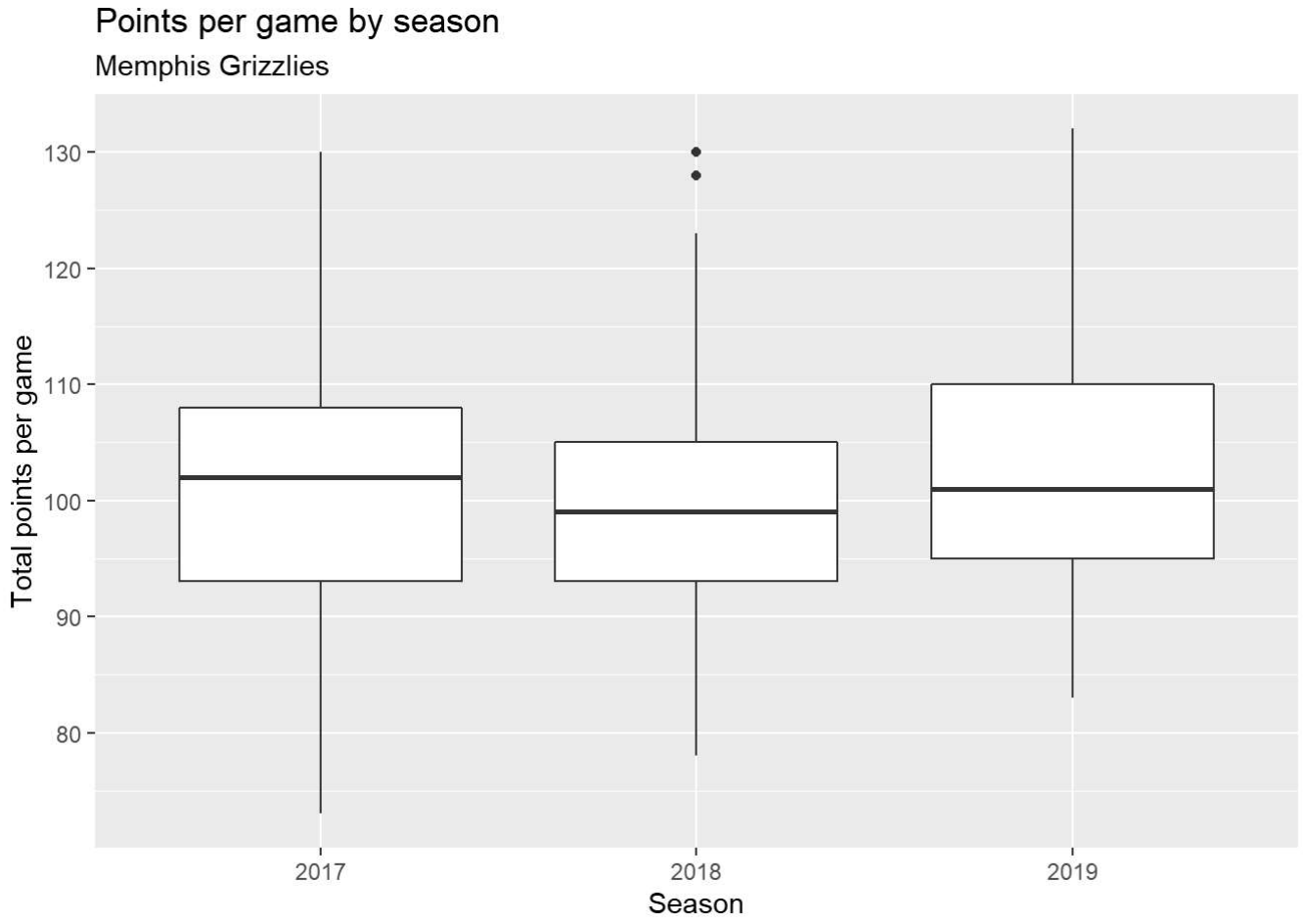
```
gms <- readRDS('../data/game_summary.Rds')

# Fewest Points
gms %>%
  group_by(nameTeam) %>%
  summarise(overallPoints = sum(pts)) %>%
  arrange(overallPoints)
```

```
## # A tibble: 30 × 2
##    nameTeam          overallPoints
##    <chr>                     <dbl>
##  1 Memphis Grizzlies         24874
##  2 Dallas Mavericks          25346
##  3 Chicago Bulls             25480
##  4 Orlando Magic             25567
##  5 Detroit Pistons           25596
##  6 Miami Heat                25608
##  7 New York Knicks           25697
##  8 Sacramento Kings          25897
##  9 Utah Jazz                 25959
## 10 Indiana Pacers            26131
## # … with 20 more rows
```

```
gms %>%
  group_by(nameTeam) %>%
  summarise(overallPoints = sum(pts),
            winPct = mean(isWin)) %>%
  arrange(winPct)
```

```
## # A tibble: 30 × 3
##    nameTeam           overallPoints winPct
##    <chr>                      <dbl>  <dbl>
##  1 Phoenix Suns               26168  0.260
##  2 New York Knicks            25697  0.313
##  3 Brooklyn Nets              26618  0.366
##  4 Chicago Bulls              25480  0.366
##  5 Dallas Mavericks           25346  0.366
##  6 Atlanta Hawks              26228  0.390
##  7 Orlando Magic              25567  0.390
##  8 Los Angeles Lakers         26602  0.398
##  9 Memphis Grizzlies          24874  0.398
## 10 Sacramento Kings           25897  0.398
## # … with 20 more rows
```

```
gms %>%
  filter(nameTeam == 'Memphis Grizzlies') %>%
  ggplot(aes(x = factor(yearSeason),y = pts)) +
  geom_boxplot() +
  labs(title = 'Points per game by season',
       subtitle = "Memphis Grizzlies",
       x = 'Season',
       y = 'Total points per game')
```



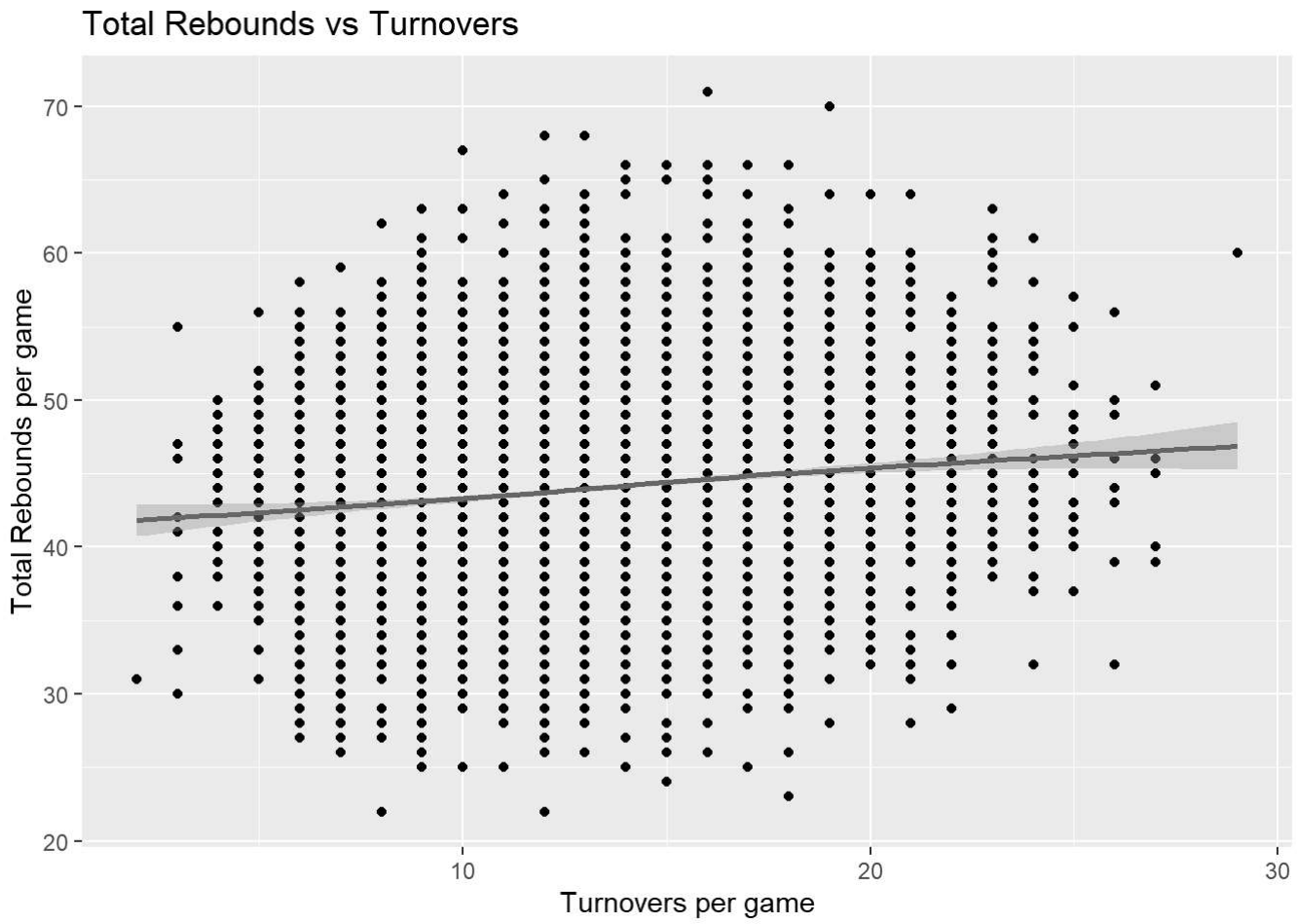
- The team with the fewest points over all three seasons is the Memphis Grizzlies. This is NOT the same as the team with the lowest win percentage, which is the Phoenix Suns who had roughly 1,500 more points than the Grizzlies. The Grizzlies were tied for 8th from the bottom in terms of win percentage.

Question 7 [5 Points]

Investigate the following theory: “Good teams rebound more and turn over the ball less. Therefore, turnovers and rebounds should be negatively correlated.” Plot the conditional relationship and either refute or support the theory. EXTRA CREDIT: provide a theorized explanation for the relationship you observe in the data.

```
gms %>%
  ggplot(aes(x = tov,y = treb)) +
  geom_point() +
  geom_smooth() +
  labs(title = 'Total Rebounds vs Turnovers',
       x = 'Turnovers per game',
       y = 'Total Rebounds per game')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

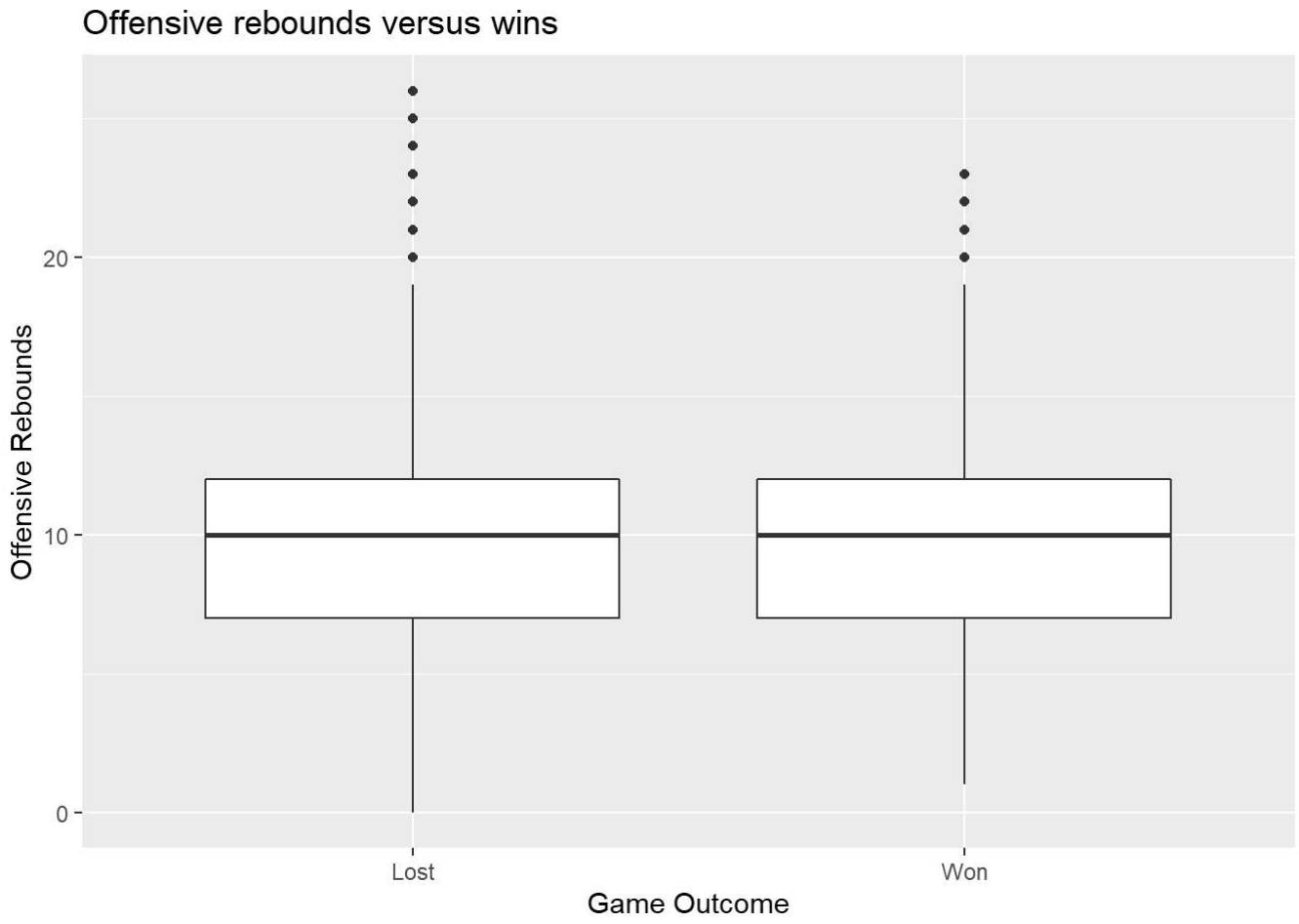


- There is no evidence to support the theory that turnovers and rebounds are negatively correlated. If anything, these two measures are *positively* correlated. This might reflect the fact teams which turnover the ball more are playing a faster style of game, which means that they also get more rebounds.

Question 8 [5 points]

Investigate the following theory: “Offensive rebounds give the team extra chances to score, more free throw opportunities, and tires out the defense. As such, offensive rebounds should be positively correlated with wins.” Plot the conditional relationship, and then use **bootstrapping** to state your confidence in your conclusion. NB: `set.seed(123)` for replicability, and calculate 100 bootstrapped simulations.

```
set.seed(123)
# Conditional relationship plot
gms %>%
  ggplot(aes(x = isWin,y = oreb)) +
  geom_boxplot() +
  labs(title = 'Offensive rebounds versus wins',
       x = 'Game Outcome',
       y = 'Offensive Rebounds') +
  scale_x_discrete(labels = c('Lost','Won'))
```



```
# Bootstraps
bsRes <- NULL
for(i in 1:100) {
  bsRes <- gms %>%
    sample_n(size = nrow(.),replace = T) %>%
    group_by(isWin) %>%
    summarise(avg_oreb = mean(oreb,na.rm=T),.groups = 'drop') %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}

bsRes %>%
  spread(isWin,avg_oreb,sep = '_') %>%
  mutate(diff = isWin_TRUE - isWin_FALSE) %>%
  summarise(mean(diff > 0))
```

```
## # A tibble: 1 × 1
##   `mean(diff > 0)`
##              <dbl>
## 1             0.09
```

```
# Alternative choice of bootstrap size
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  bsRes <- gms %>%
    sample_n(size = 1000,replace = T) %>%
    group_by(isWin) %>%
    summarise(avg_oreb = mean(oreb,na.rm=T),.groups = 'drop') %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}

bsRes %>%
  spread(isWin,avg_oreb,sep = '_') %>%
  mutate(diff = isWin_TRUE - isWin_FALSE) %>%
  summarise(mean(diff > 0))
```

```
## # A tibble: 1 × 1
##   `mean(diff > 0)`
##              <dbl>
## 1             0.36
```

- There is no evidence to support the theory that offensive rebounds lead to wins. The descriptive plot finds no visual difference in the offensive wins for teams that win or lose. Furthermore, out of 100 bootstrapped simulations, winning teams had more offensive rebounds than losing teams in only 9 of the simulations. If anything, these results suggest that offensive rebounds are associated with *losing*.

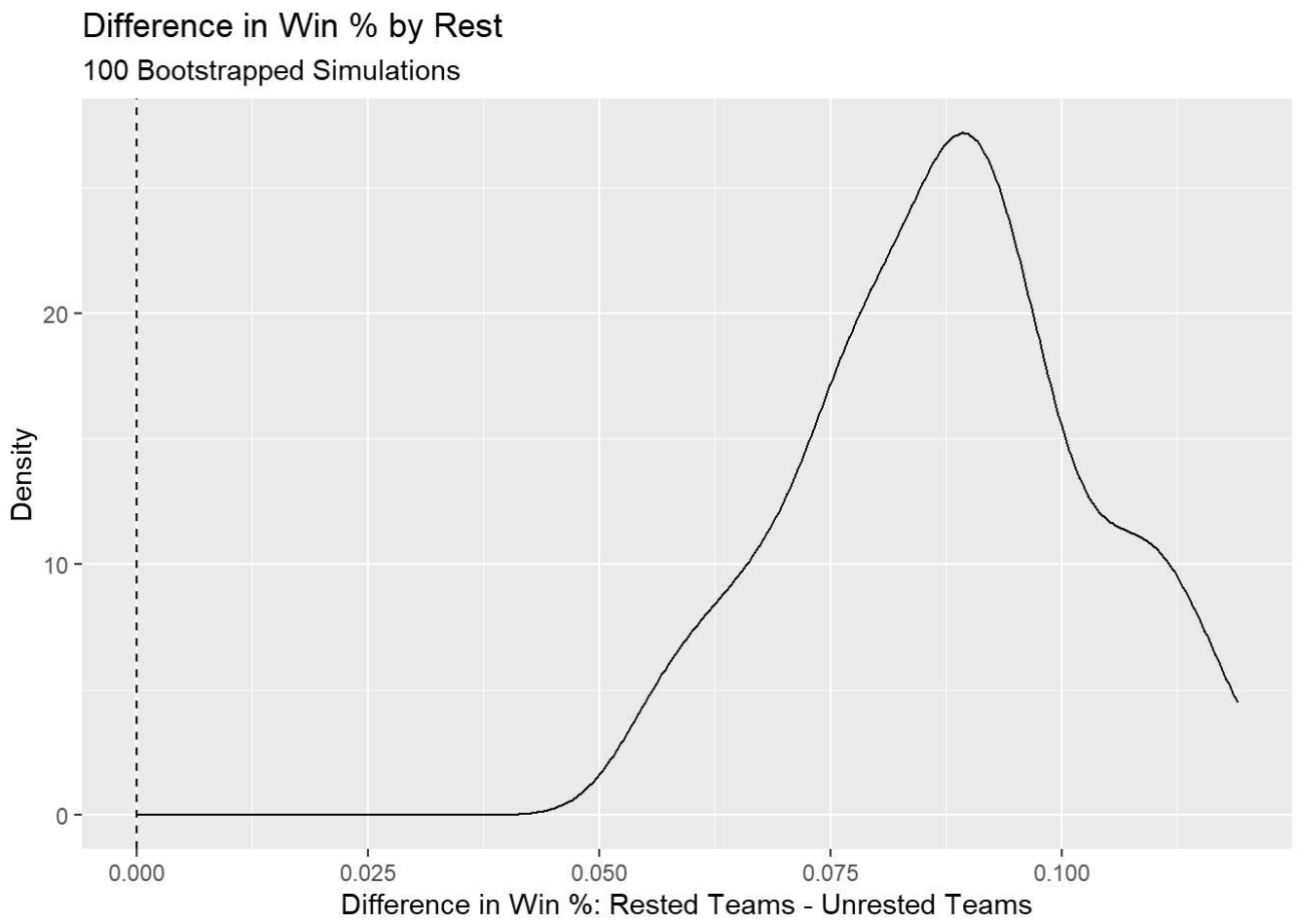
Question 9 [5 points]

Test the following hypothesis: “Teams with at least 1 day rest between games are more likely to win their next game than those with no rest.” You will need to create a new variable that is a dummy with the value of 1 if `teamrest` is greater than 0, and zero otherwise. Use bootstrapping to express your **confidence** about your conclusion. Plot the difference in win probabilities across 100 bootstrapped simulations. NB: `set.seed(123)` for replicability.

```
set.seed(123)

# Overall
gms %>%
  mutate(rest = ifelse(teamrest > 0,1,0)) %>%
  group_by(rest) %>%
  summarise(propWin = mean(isWin))
```

```
## # A tibble: 2 × 2
##    rest propWin
##   <dbl>   <dbl>
## 1     0   0.430
## 2     1   0.515
```

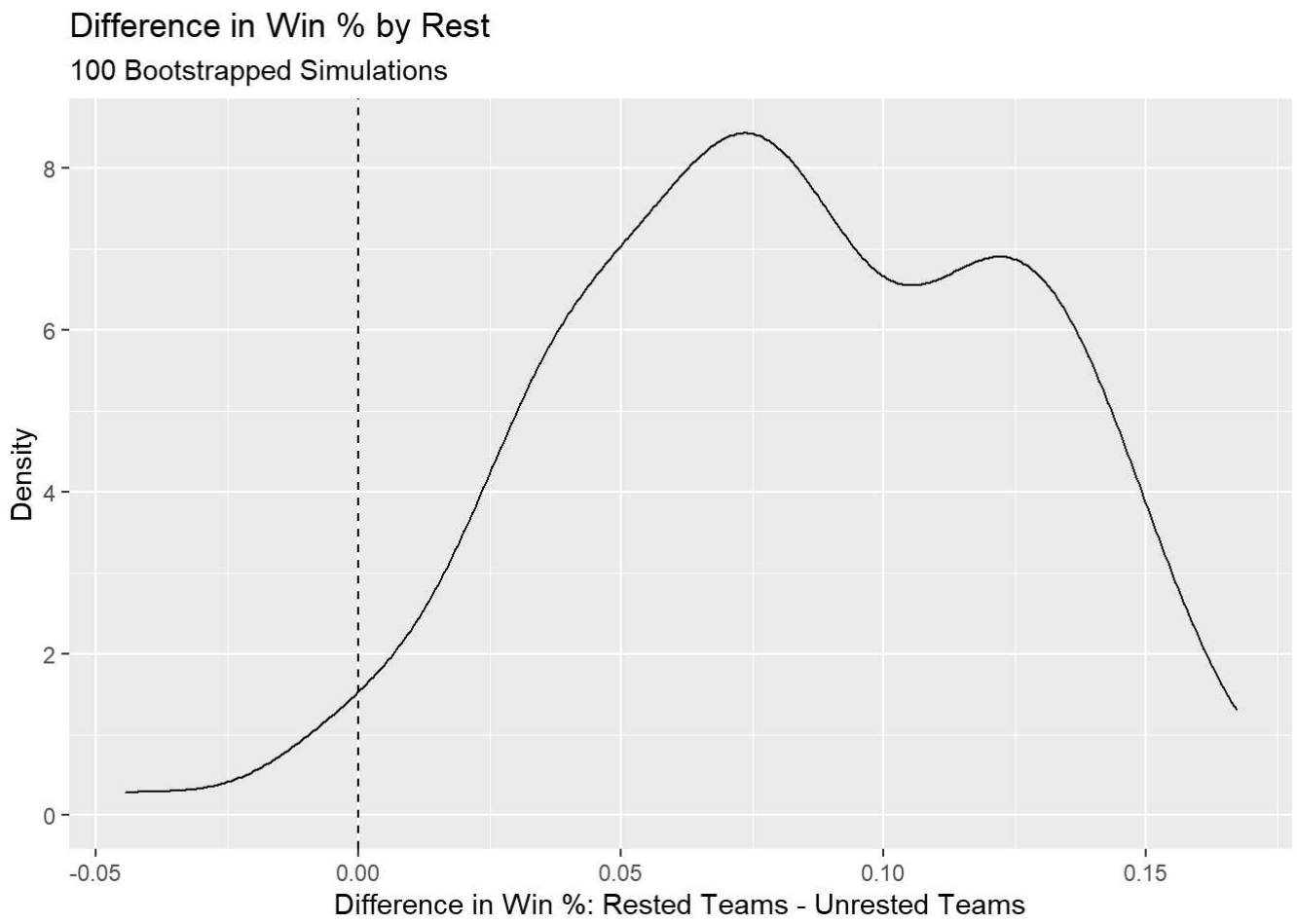
```
# Bootstraps
bsRes <- NULL
for(i in 1:100) {
  bsRes <- gms %>%
    mutate(rest = ifelse(teamrest > 0,1,0)) %>%
    sample_n(size = nrow(.),replace = T) %>%
    group_by(rest) %>%
    summarise(propWin = mean(isWin)) %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}


toplot <- bsRes %>%
  spread(rest,propWin,sep = '_') %>%
  mutate(winDiff = rest_1 - rest_0)

# Plot
toplot %>%
  ggplot(aes(x = winDiff)) +
  geom_density() +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  labs(title = "Difference in Win % by Rest",
       subtitle = "100 Bootstrapped Simulations",
       x = "Difference in Win %: Rested Teams - Unrested Teams",
       y = "Density")
```




```
# Confidence
toplot %>%
  summarise(mean(winDiff > 0))
```

```
## # A tibble: 1 × 1
##   `mean(winDiff > 0)`
##                 <dbl>
## 1                   1
```

```
# Alternative choice of bootstrap size
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  bsRes <- gms %>%
    mutate(rest = ifelse(teamrest > 0,1,0)) %>%
    sample_n(size = 1000,replace = T) %>%
    group_by(rest) %>%
    summarise(propWin = mean(isWin)) %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}


toplot <- bsRes %>%
  spread(rest,propWin,sep = '_') %>%
  mutate(winDiff = rest_1 - rest_0)

# Plot
toplot %>%
  ggplot(aes(x = winDiff)) +
  geom_density() +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  labs(title = "Difference in Win % by Rest",
       subtitle = "100 Bootstrapped Simulations",
       x = "Difference in Win %: Rested Teams - Unrested Teams",
       y = "Density")
```




```
# Confidence
toplot %>%
  summarise(mean(winDiff > 0))
```

```
## # A tibble: 1 × 1
##   `mean(winDiff > 0)`
##                 <dbl>
## 1                0.97
```

- Teams are more likely to win when they have 1 or more days off between games. Overall, rested teams have a 51.5% win percentage while unrested teams have a 43% win percentage. Rested teams have a higher win percentage than unrested teams in 97% of bootstrapped simulations.

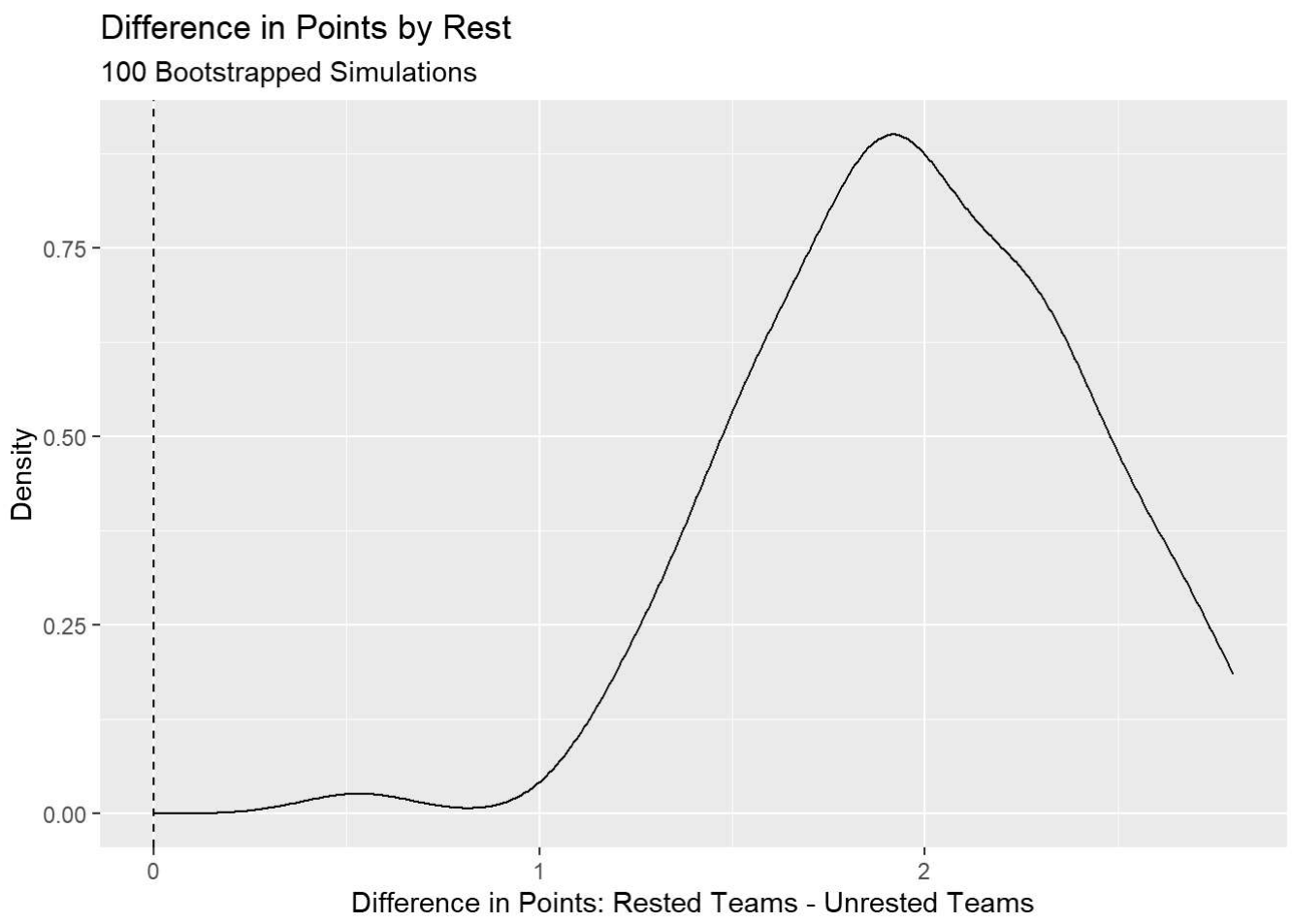
Question 10 [5 points]

Run the same analysis for points (`pts`). How certain are you that teams with 1+ days of rest score more points than those without?

```
# Points
set.seed(123)

# Overall
gms %>%
  mutate(rest = ifelse(teamrest > 0,1,0)) %>%
  group_by(rest) %>%
  summarise(avgPts = mean(pts))
```

```
## # A tibble: 2 × 2
##    rest avgPts
##   <dbl>  <dbl>
## 1     0   106.
## 2     1   108.
```

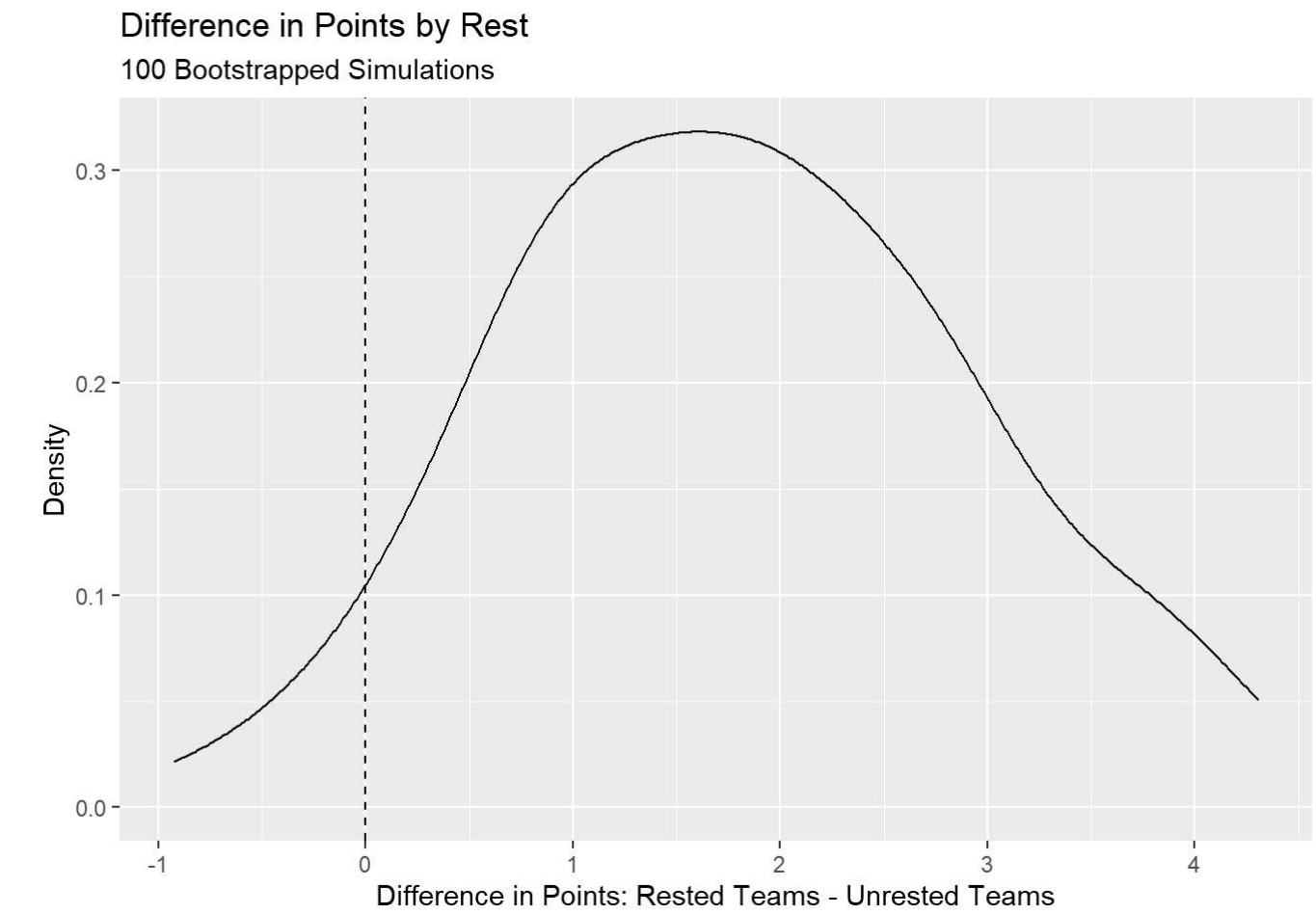
```
# Bootstraps
bsRes <- NULL
for(i in 1:100) {
  bsRes <- gms %>%
    sample_n(size = nrow(.),replace = T) %>%
    mutate(rest = ifelse(teamrest > 0,1,0)) %>%
    group_by(rest) %>%
    summarise(avgPts = mean(pts)) %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}


toplot <- bsRes %>%
  spread(rest,avgPts,sep = '_') %>%
  mutate(ptsDiff = rest_1 - rest_0)

# Plot
toplot %>%
  ggplot(aes(x = ptsDiff)) +
  geom_density() +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  labs(title = "Difference in Points by Rest",
       subtitle = "100 Bootstrapped Simulations",
       x = "Difference in Points: Rested Teams - Unrested Teams",
       y = "Density")
```




```
# Confidence
toplot %>%
  summarise(mean(ptsDiff > 0))
```

```
## # A tibble: 1 × 1
##   `mean(ptsDiff > 0)`
##                 <dbl>
## 1                   1
```

```
# Alternative choice of bootstrap size
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  bsRes <- gms %>%
    sample_n(size = 1000,replace = T) %>%
    mutate(rest = ifelse(teamrest > 0,1,0)) %>%
    group_by(rest) %>%
    summarise(avgPts = mean(pts)) %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}


toplot <- bsRes %>%
  spread(rest,avgPts,sep = '_') %>%
  mutate(ptsDiff = rest_1 - rest_0)

# Plot
toplot %>%
  ggplot(aes(x = ptsDiff)) +
  geom_density() +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  labs(title = "Difference in Points by Rest",
       subtitle = "100 Bootstrapped Simulations",
       x = "Difference in Points: Rested Teams - Unrested Teams",
       y = "Density")
```




```
# Confidence
toplot %>%
  summarise(mean(ptsDiff > 0))
```

```
## # A tibble: 1 × 1
##   `mean(ptsDiff > 0)`
##                 <dbl>
## 1                0.96
```

- Rested teams score an average of two more points per game than those teams without rest. Rested teams score more points than unrested teams in 96 out of 100 bootstrapped simulations.

Question 11 [10 Extra Credit Points]

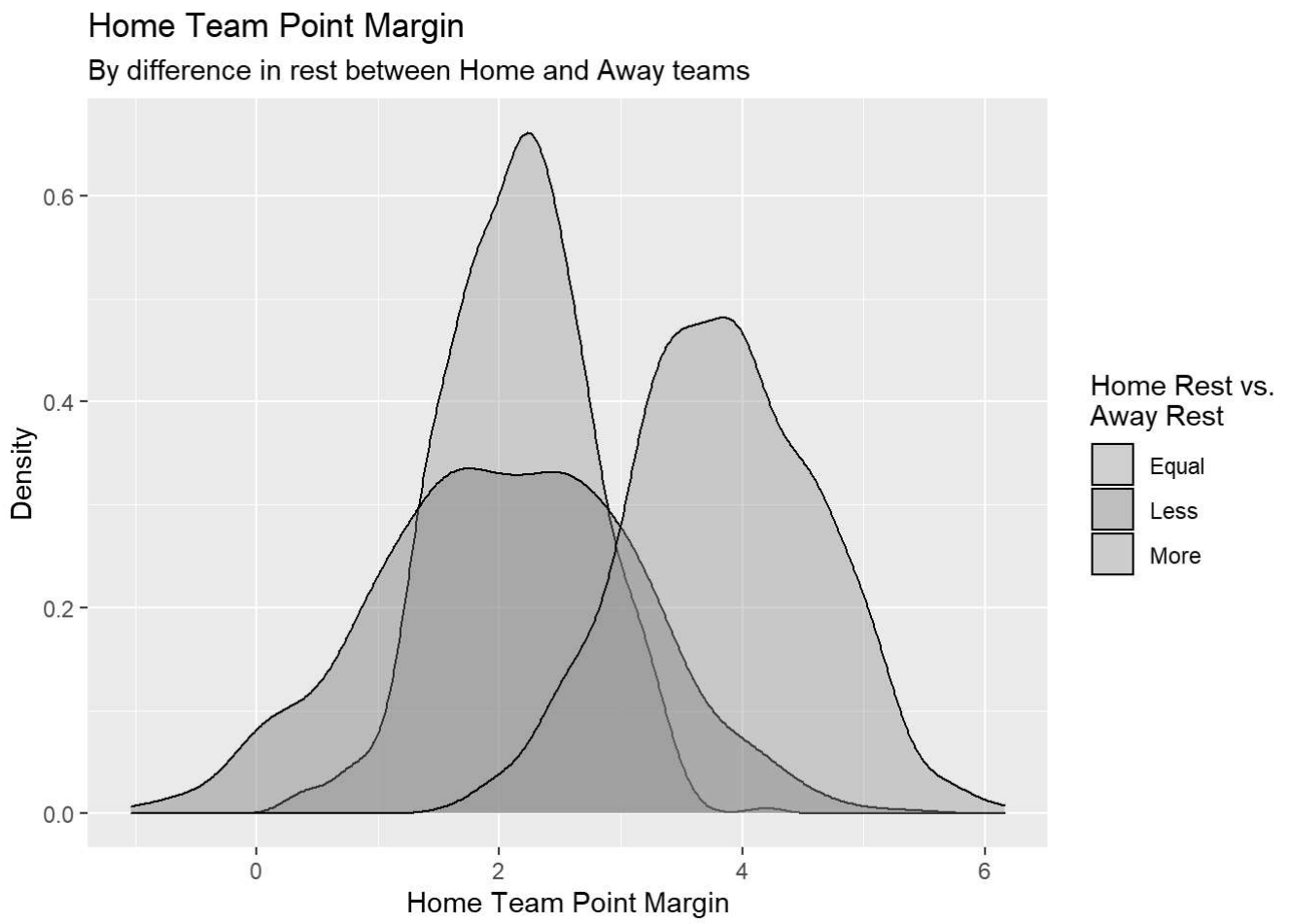
EXTRA CREDIT: Is the point difference closer when both teams are equally well-rested? To answer this question, you will need to calculate the rest difference between the two teams per `idGame`, as well as their point difference. Plot the distribution of point margins for less rested and more rested teams separately, then calculate the difference between less and more rested teams and plot this. Express your conclusion in terms of certainty using 500 bootstrapped simulations.

Hint: select just the game ID, the home/away status, and either the rest or the points. Then use `spread(locationGame,teamrest)` or `spread(locationGame,pts)` to get two columns of rest and points for the home and away teams. Calculate the rest and points difference consistently (i.e., always H-A or A-H), select just the idGame and difference measure, and then combine to create a new dataset via `left_join`. Armed with these data, conduct a bootstrapped test of the conditional relationship between rest differences and point differences.

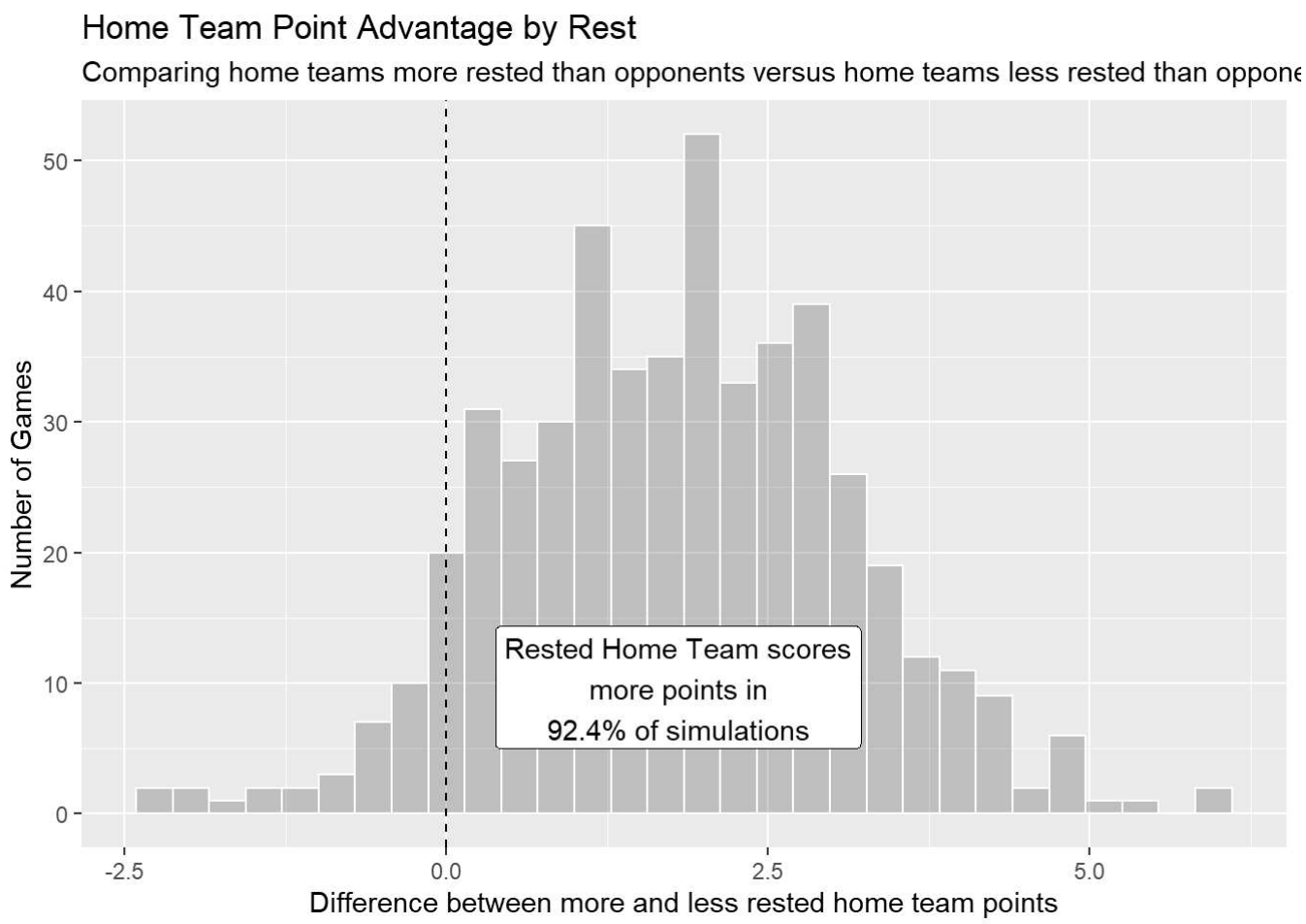
```
set.seed(123)
# Wrangling
restDiff <- gms %>%
  select(idGame,locationGame,teamrest) %>%
  spread(locationGame,teamrest) %>%
  mutate(restDiff = H - A) %>%
  select(idGame,restDiff)

pointDiff <- gms %>%
  select(idGame,locationGame,pts) %>%
  spread(locationGame,pts) %>%
  mutate(ptsDiff = H - A) %>%
  select(idGame,ptsDiff)

diffDat <- restDiff %>%
  filter(abs(restDiff) < 10) %>%
  left_join(pointDiff) %>%
  mutate(restDiff = ifelse(restDiff < 0,'Less',
                           ifelse(restDiff > 0,'More','Equal')))
```

```
## Joining, by = "idGame"
```

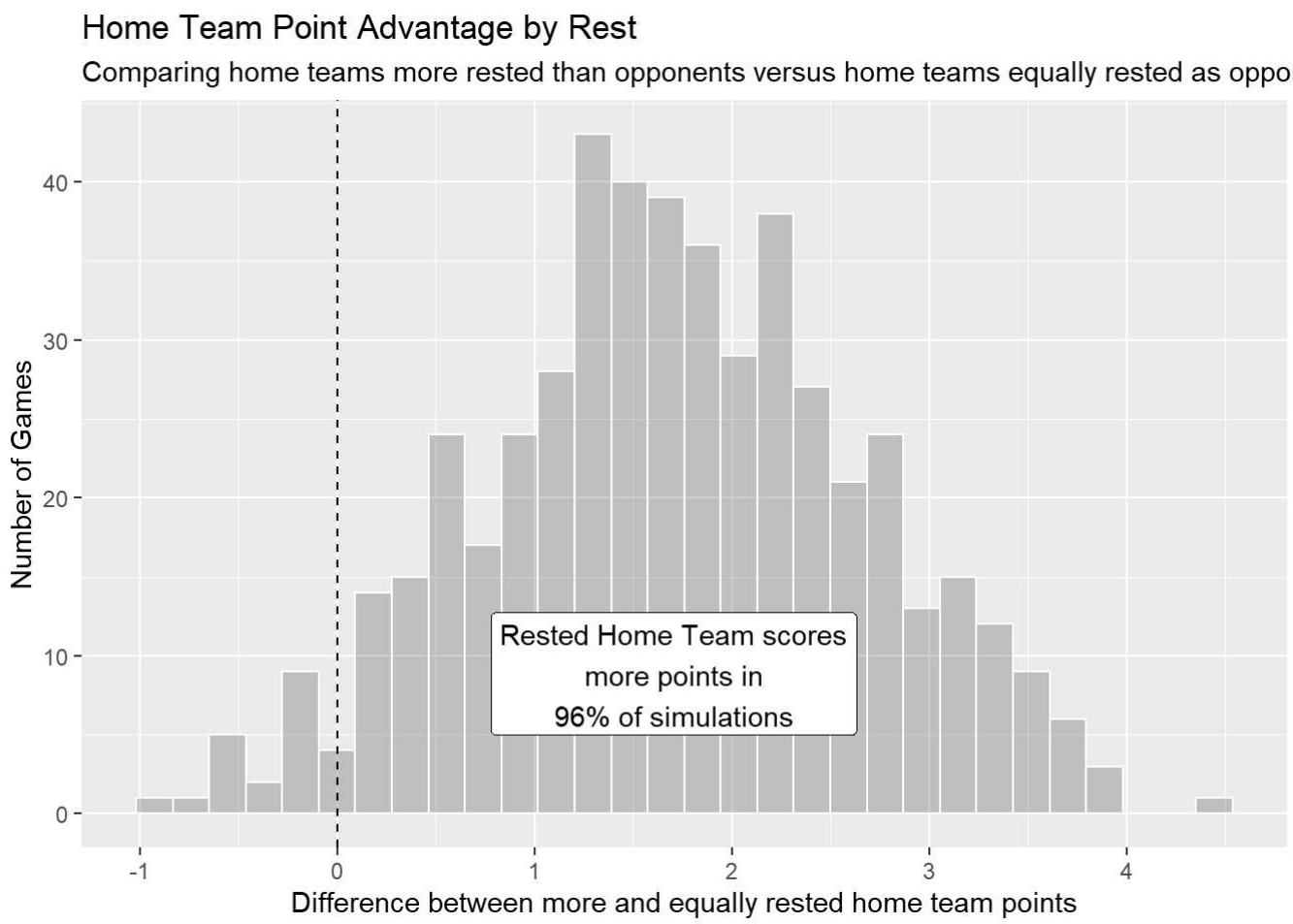
```
# Bootstrapping
bsRes <- NULL
for(i in 1:500) {
  bsRes <- diffDat %>%
    sample_n(1000,replace = T) %>%
    group_by(restDiff) %>%
    summarise(ptsDiff = mean(ptsDiff)) %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}

# Raw Plot by Rest
bsRes %>%
  ggplot(aes(x = ptsDiff,fill = restDiff)) +
  geom_density(alpha = .4,color = 'black') +
  labs(title = "Home Team Point Margin",
       subtitle = "By difference in rest between Home and Away teams",
       x = "Home Team Point Margin",
       y = "Density",
       fill = "Home Rest vs.\nAway Rest")
```



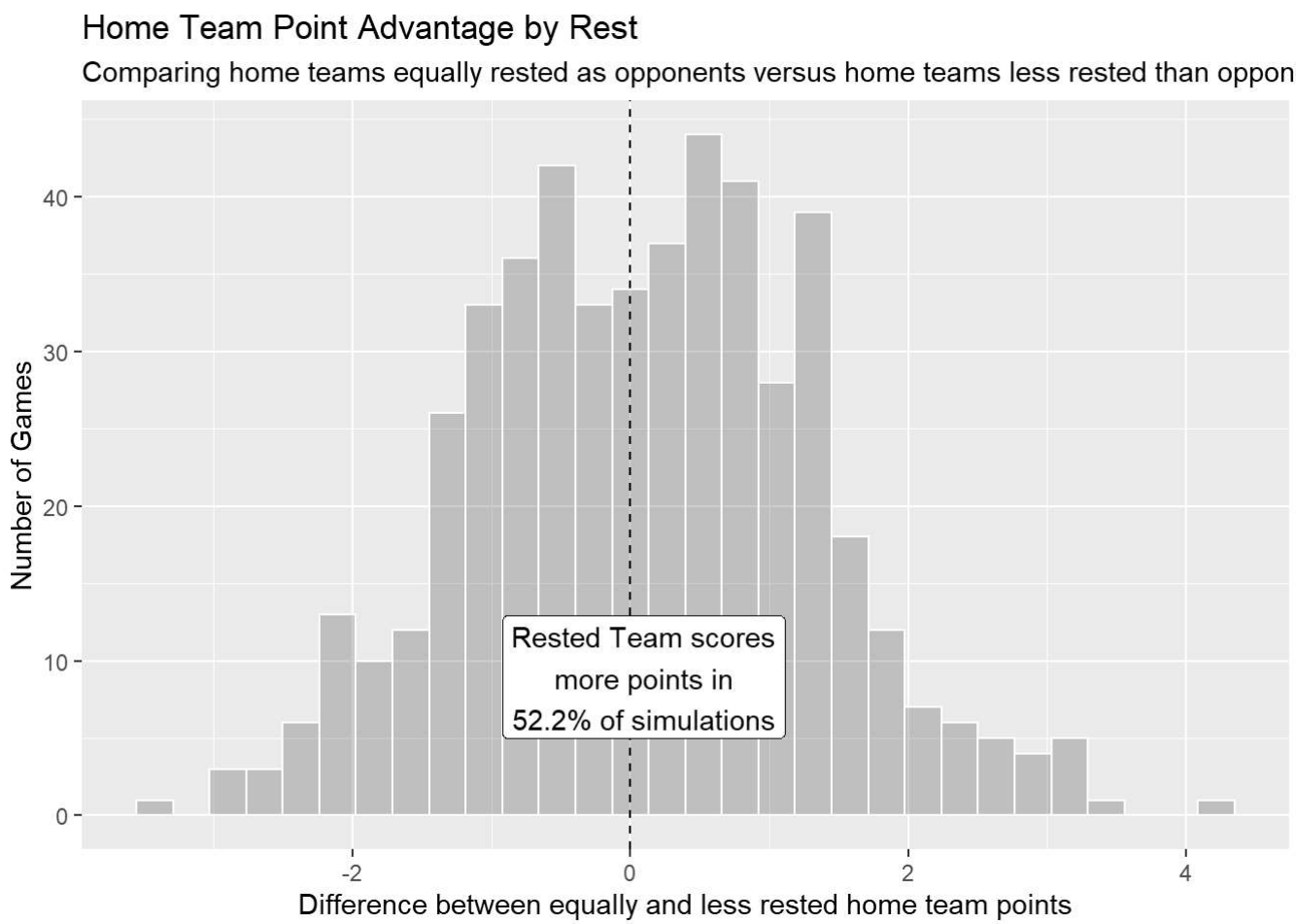

```
# Difference 1 plot: More versus Less rested
toplotMvL <- bsRes %>%
  spread(restDiff,ptsDiff) %>%
  mutate(diff = More - Less)

toplotMvL %>%
  ggplot(aes(x = diff)) +
  geom_histogram(alpha = .3,color = 'white',bins = 30) +
  labs(title = "Home Team Point Advantage by Rest",
       subtitle = "Comparing home teams more rested than opponents versus home teams less rested
than opponents",
       y = "Number of Games",
       x = "Difference between more and less rested home team points") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  annotate(geom = 'label',x = mean(toplotMvL$diff),y = 5,vjust = 0,
           label = paste0('Rested Home Team scores\nmore points in\n',round(mean(toplotMvL$diff>
0),3)*100,'% of simulations'))
```




```
# Difference 2 plot: More versus Equally rested
toplotMvE <- bsRes %>%
  spread(restDiff,ptsDiff) %>%
  mutate(diff = More - Equal)

toplotMvE %>%
  ggplot(aes(x = diff)) +
  geom_histogram(alpha = .3,color = 'white',bins = 30) +
  labs(title = "Home Team Point Advantage by Rest",
       subtitle = "Comparing home teams more rested than opponents versus home teams equally res
ted as opponents",
       y = "Number of Games",
       x = "Difference between more and equally rested home team points") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  annotate(geom = 'label',x = mean(toplotMvE$diff),y = 5,vjust = 0,
           label = paste0('Rested Home Team scores\nmore points in\n',round(mean(toplotMvE$diff>
0),3)*100,'% of simulations'))
```




```
# Difference 3 Plot: Equally versus Less rested
toplotEvL <- bsRes %>%
  spread(restDiff,ptsDiff) %>%
  mutate(diff = Equal - Less)

toplotEvL %>%
  ggplot(aes(x = diff)) +
  geom_histogram(alpha = .3,color = 'white',bins = 30) +
  labs(title = "Home Team Point Advantage by Rest",
       subtitle = "Comparing home teams equally rested as opponents versus home teams less reste
d than opponents",
       y = "Number of Games",
       x = "Difference between equally and less rested home team points") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  annotate(geom = 'label',x = mean(toplotEvL$diff),y = 5,vjust = 0,
           label = paste0('Rested Team scores\nmore points in\n',round(mean(toplotEvL$diff>0),3)
*100,'% of simulations'))
```



- If the home team has more rest than their opponent, they score more points than a home team that is less rested than their opponent (96% confidence). If the home team has more rest than their opponent, they score more points than a home team that is equally rested as their opponent (96% confidence). If the home team is equally rested as their opponent, they don't score more points than a home team that is less rested than their opponent (52.2% confidence). Overall, home teams that are more rested than their opponent score 4 more points than their opponent, while home teams that are equally rested as their opponent and home teams that are less rested than their opponent score roughly 2 more points than their opponent.

Optional Extra Credit [5 Extra Credit Points]

Please complete this **anonymous** course evaluation. This does not influence Professor Bisbee's career or position in the university and will only be used to improve the course. You can find the anonymous survey here (https://nyu.qualtrics.com/jfe/form/SV_daPytGPjypTqJgy). Upon completing the survey, you will be given a completion code, which you should paste back at the end of your midterm below.

NOTE: There is only one completion code to ensure that all responses are anonymized and can't be linked back to the midterm exams. To prevent students from sharing the code with their friends to get the 5 extra credit points without completing the survey, these 5 points are only provided if the number of midterms with the completion

code *exactly equals the number of survey responses*. In other words, if there are 150 exams with the completion code, but only 50 completed surveys, **all students will forfeit their extra credit points**. The purpose of this strict rule is to disincentivize the sharing of this code either by those who would fill out the survey and then share the code, or by those who would ask to be given the code without filling out the survey.

- `D@taSc!enceForEveryone`