HW3 - SDS 315

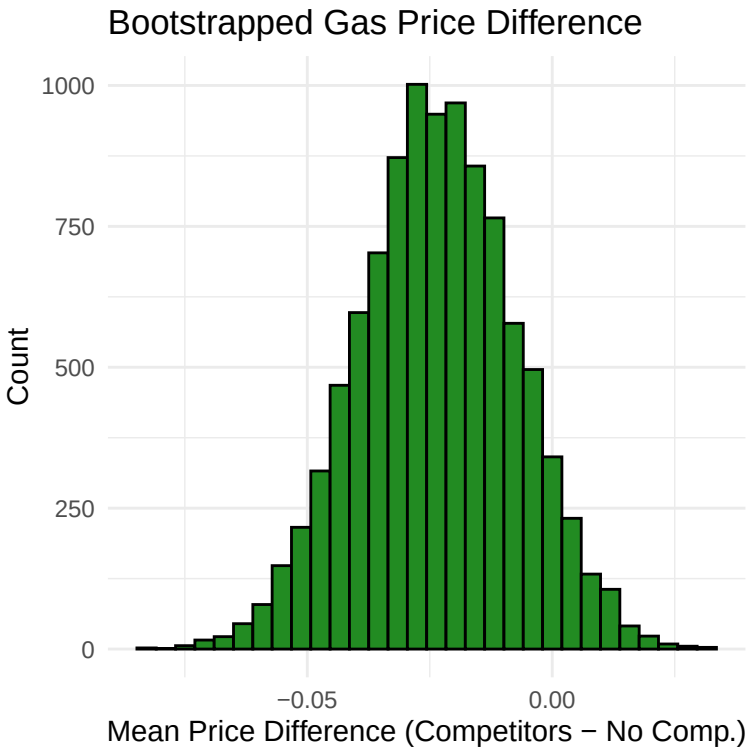
Ahantya Sharma

UT EID: as236366

Github Link: https://github.com/Ahantya/SDS315/blob/main/HW3/HW3Markdown.Rmd

Problem 1 - Gas Prices

A.



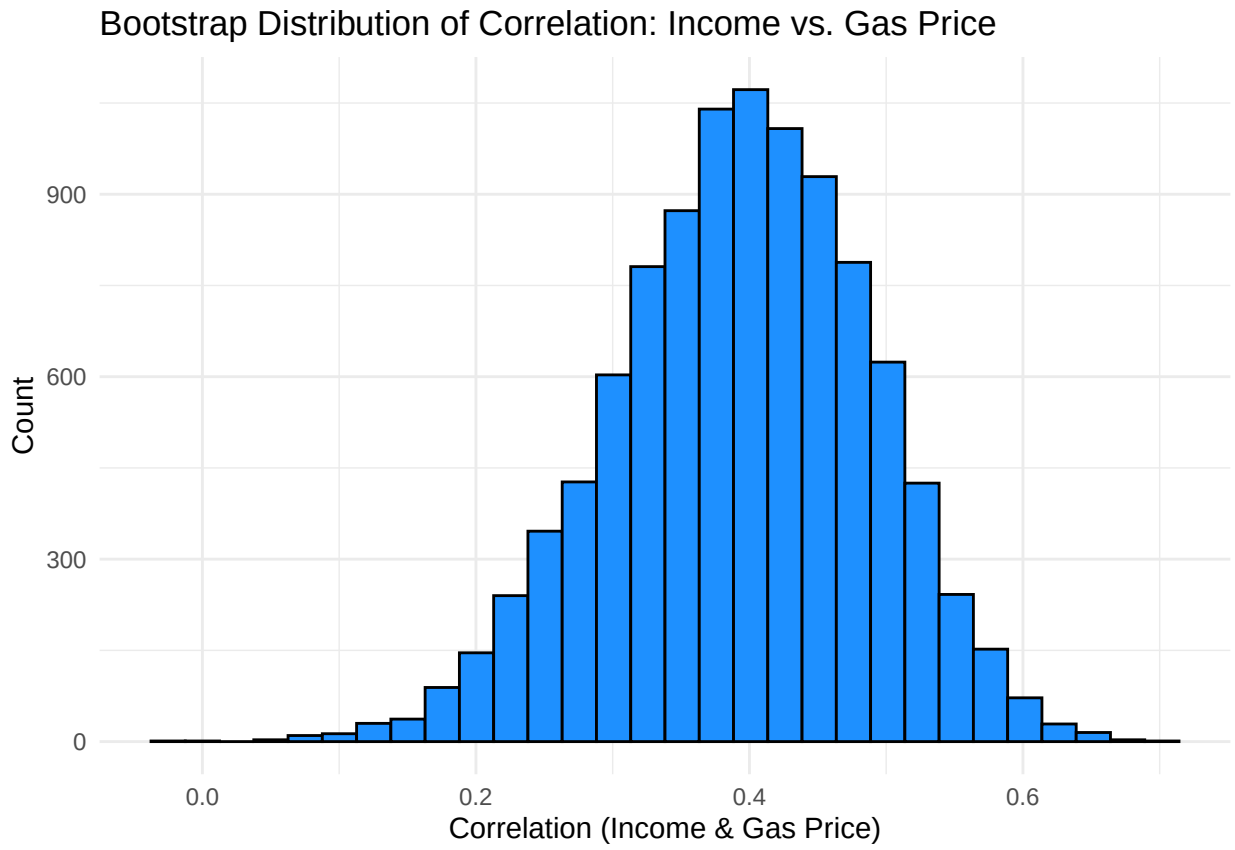

```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05479666 0.007738903  0.95 percentile -0.03164835
```

Claim: Gas stations charge more if they lack direct competition in sight.

Evidence: At a 95% confidence interval, the difference in mean prices between gas stations with competitors and gas stations without competitors is from -0.056 to 0.008 dollars. Because a possible difference of zero dollars is contained within this confidence interval, there is no statistical significance (at the 5% significance level) to claim that gas stations charge more if they lack direct competition in sight.

Conclusion: The theory is **unsupported** by the evidence, as there is no data to support a statistically signficant mean difference between prices and gas stations with / without competition in sight.

B.



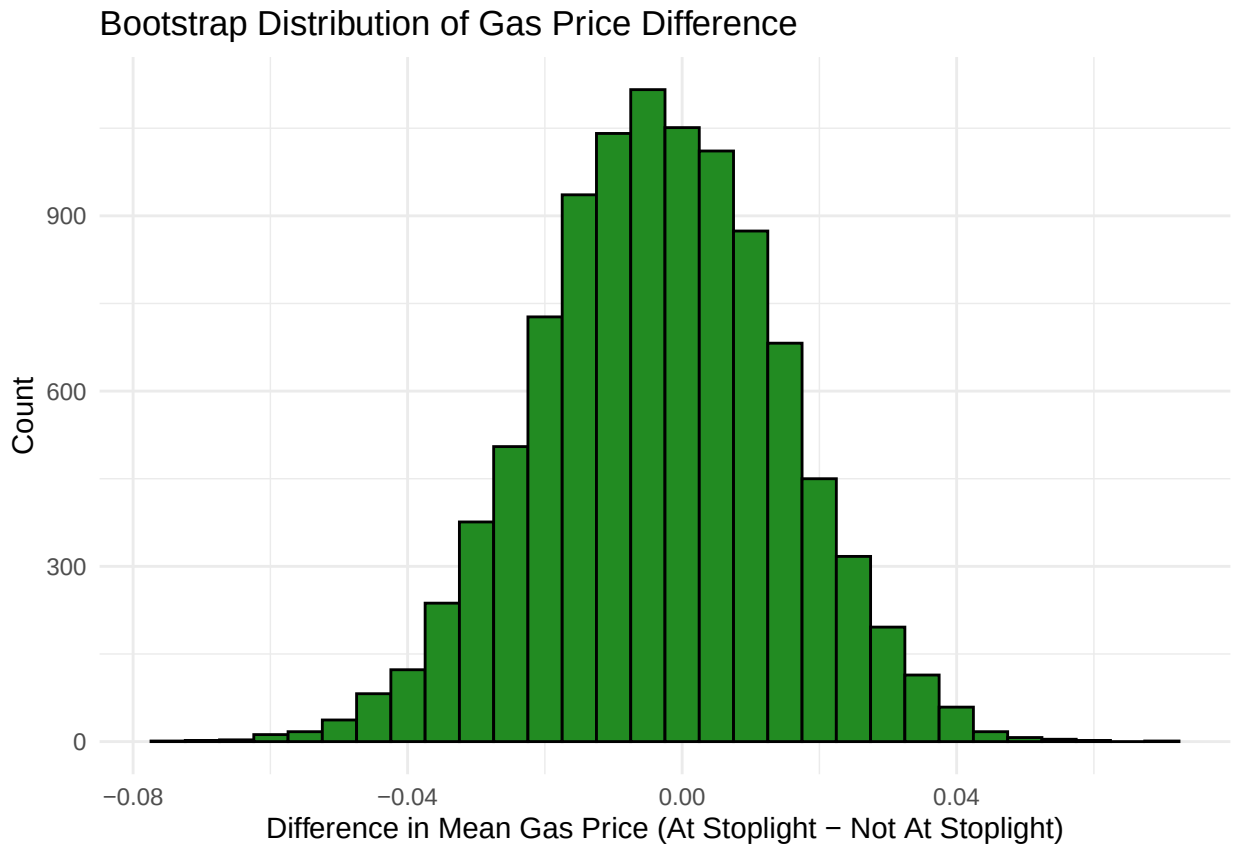

```
##     name     lower     upper level     method  estimate
## 1 result 0.2006453 0.5665537  0.95 percentile 0.5207862
```

Claim: The richer the area, the higher the gas prices.

Evidence: At a 95% confidence interval, the correlation in prices between gas stations and the median household income in the surrounding of the gas station ranges from 0.196 to 0.567 dollars. Since the possibility difference of zero dollars is not contained within the interval, there is statistical significance (at the 5% significance level) to claim that the richer the area, the higher the gas prices.

Conclusion: The theory is **supported** by the evidence, as there is data to support a statistically significant correlation between the median household income and the gas prices.

C.



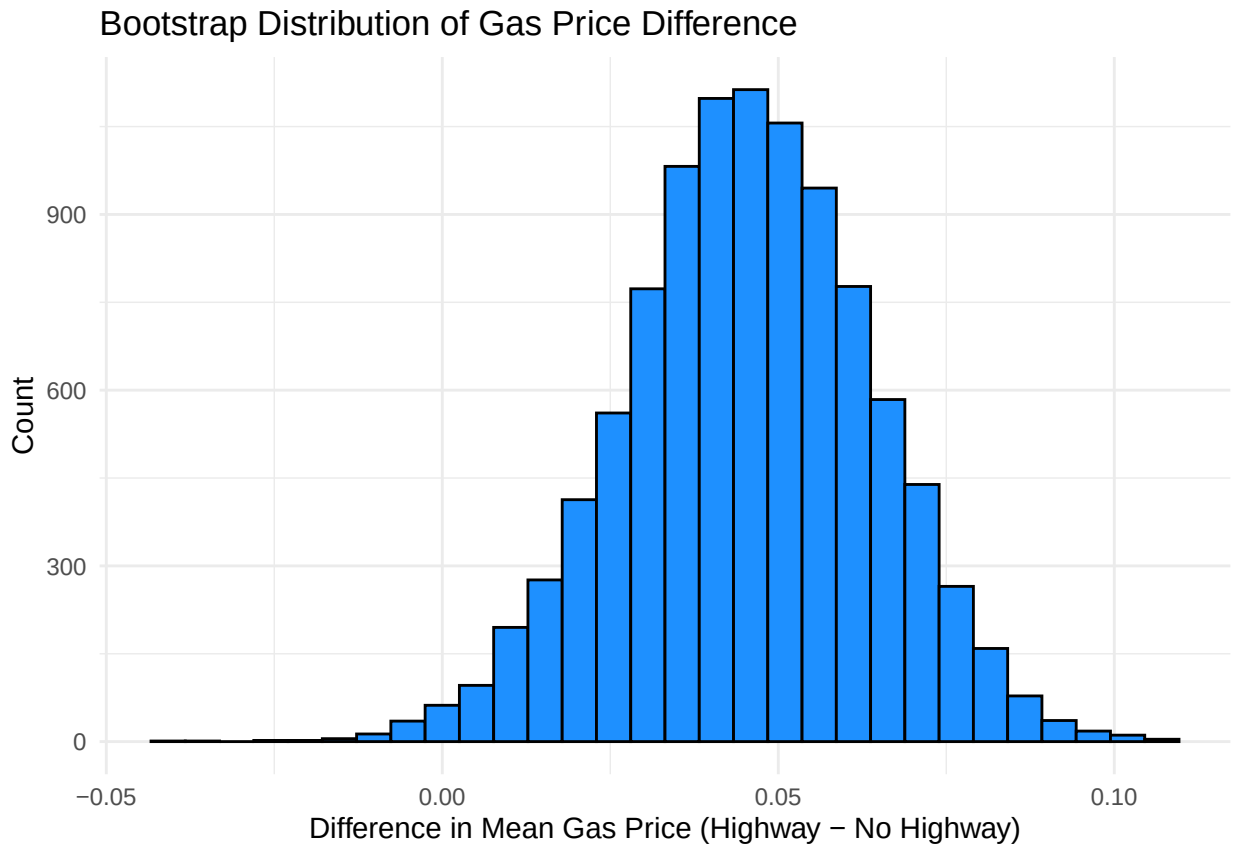

```
##       name       lower      upper level     method   estimate
## 1 diffmean -0.03818463 0.03100089  0.95 percentile -0.0143956
```

Claim: Gas stations at stoplights charge more.

Evidence: At a 95% confidence interval, the difference in mean prices for gas stations at stoplights and gas stations not at stoplights ranges from -0.039 to 0.306 dollars. Because a possible difference of zero dollars is contained within this confidence interval, there is no statistical significance (at the 5% significance level) to claim that gas stations at stoplights charge more.

Conclusion: The theory is **unsupported** by the evidence, as there is no data to support a statistically significant mean difference between gas station prices at a stoplight and gas station prices not at a stoplight.

D.



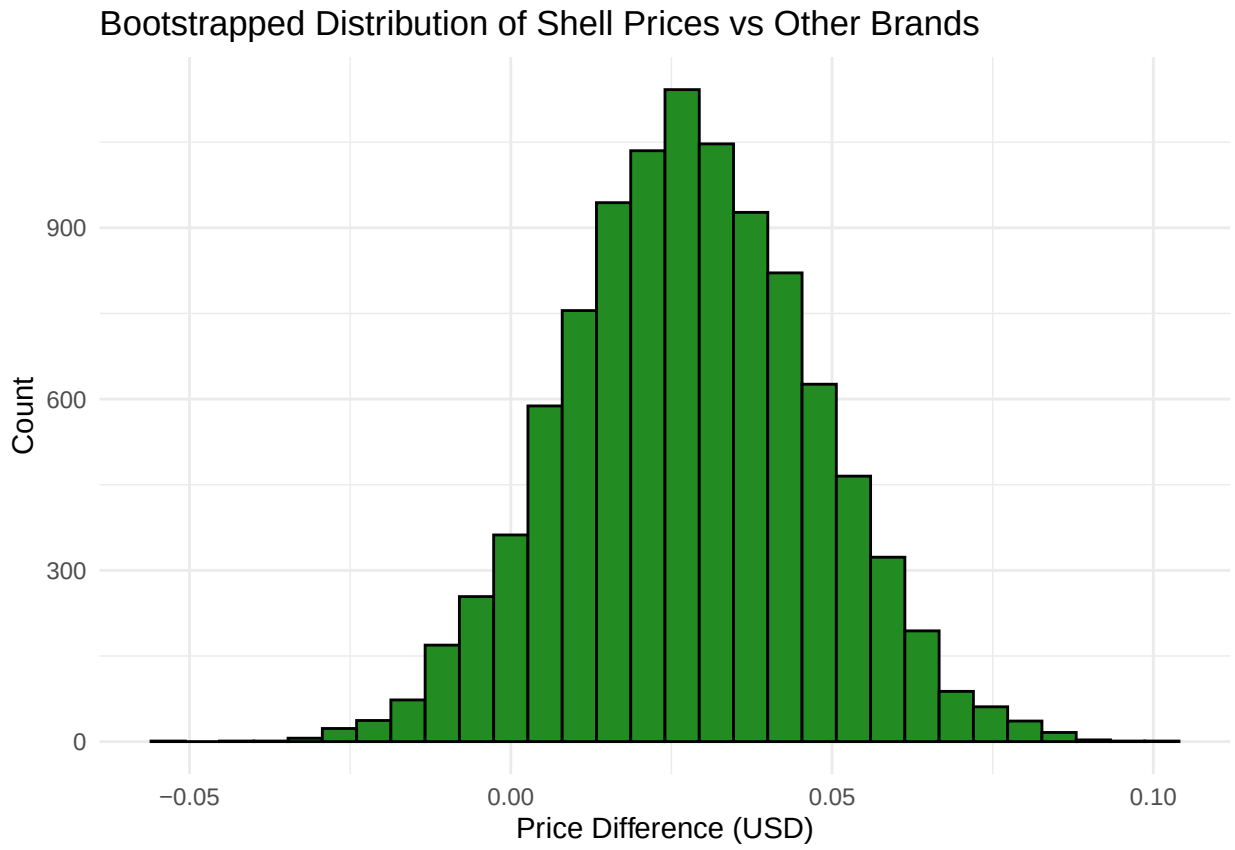

```
##       name       lower      upper level     method   estimate
## 1 diffmean 0.008612798 0.08067367  0.95 percentile 0.03519411
```

Claim: Gas stations with direct highway access charge more.

Evidence: At a 95% confidence interval, the difference in mean prices for gas stations with direct highway access and gas stations without direct highway access ranges from 0.008 to 0.082 dollars. Since the possibility difference of zero dollars is not contained within the interval, there is statistical significance (at the 5% significance level) to claim that gas stations with direct highway access charge slightly more than gas stations without direct highway access.

Conclusion: The theory is **supported** by the evidence, as there is data to support a statistically significant mean difference between gas stations with direct highway access and gas stations without direct highway access.

E.




```
##       name        lower      upper level     method   estimate
## 1 diffmean -0.009444945 0.06558139  0.95 percentile 0.03845785
```

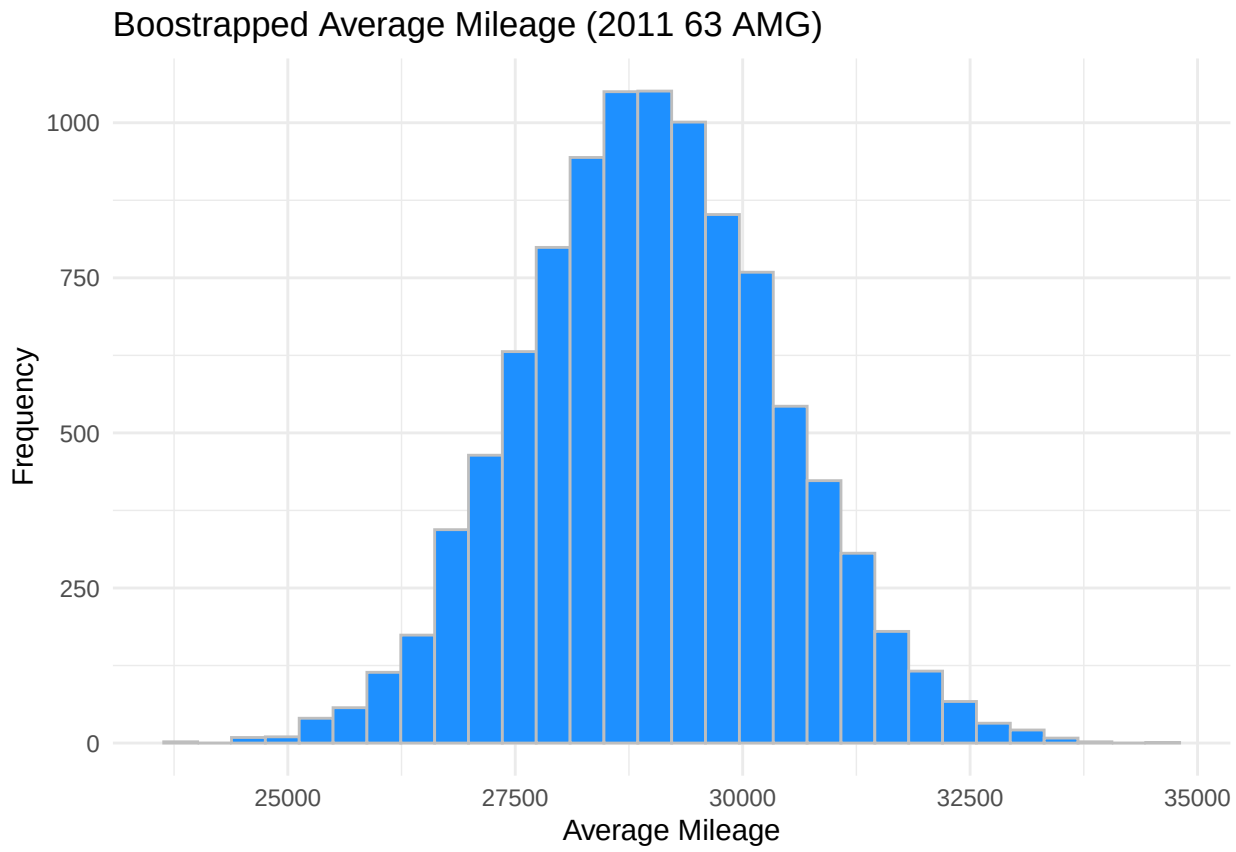
Claim: Shell charges more than all other non-Shell brands.

Evidence: At a 95% confidence interval, the mean difference in gas station prices at Shell Gas Stations and non-Shell gas stations range from -0.011 to 0.065 dollars. Because a possible difference of zero dollars is contained within this confidence interval, there is no statistical significance (at the 5% significance level) to claim that Shell charges more than all other non-Shell branded gas stations.

Conclusion: The theory is **unsupported** by the evidence, as there is no data to support a statistically significant mean difference between gas station prices at Shell gas stations and gas station prices at non-Shell gas stations

Problem 2 - Mercedes S-Class Vehicles

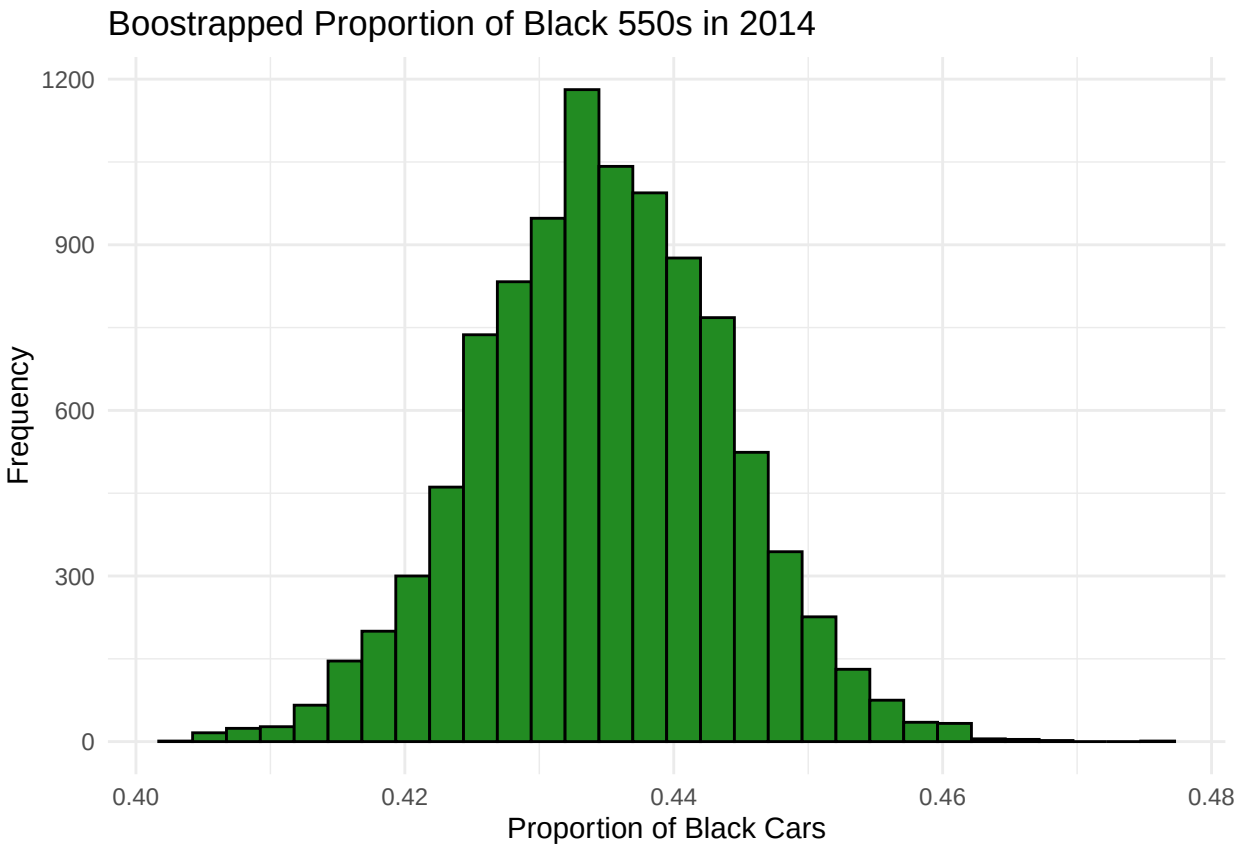
Part A - 2011 S-Class 63 AMGs




```
##   name   lower   upper level     method estimate
## 1 mean 26286.8 31806.1  0.95 percentile 28316.22
```

With 95% confidence, the average mileage of all 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected ranged from **26226 to 31823 miles**.

Part B - 2014 S-Class 550s

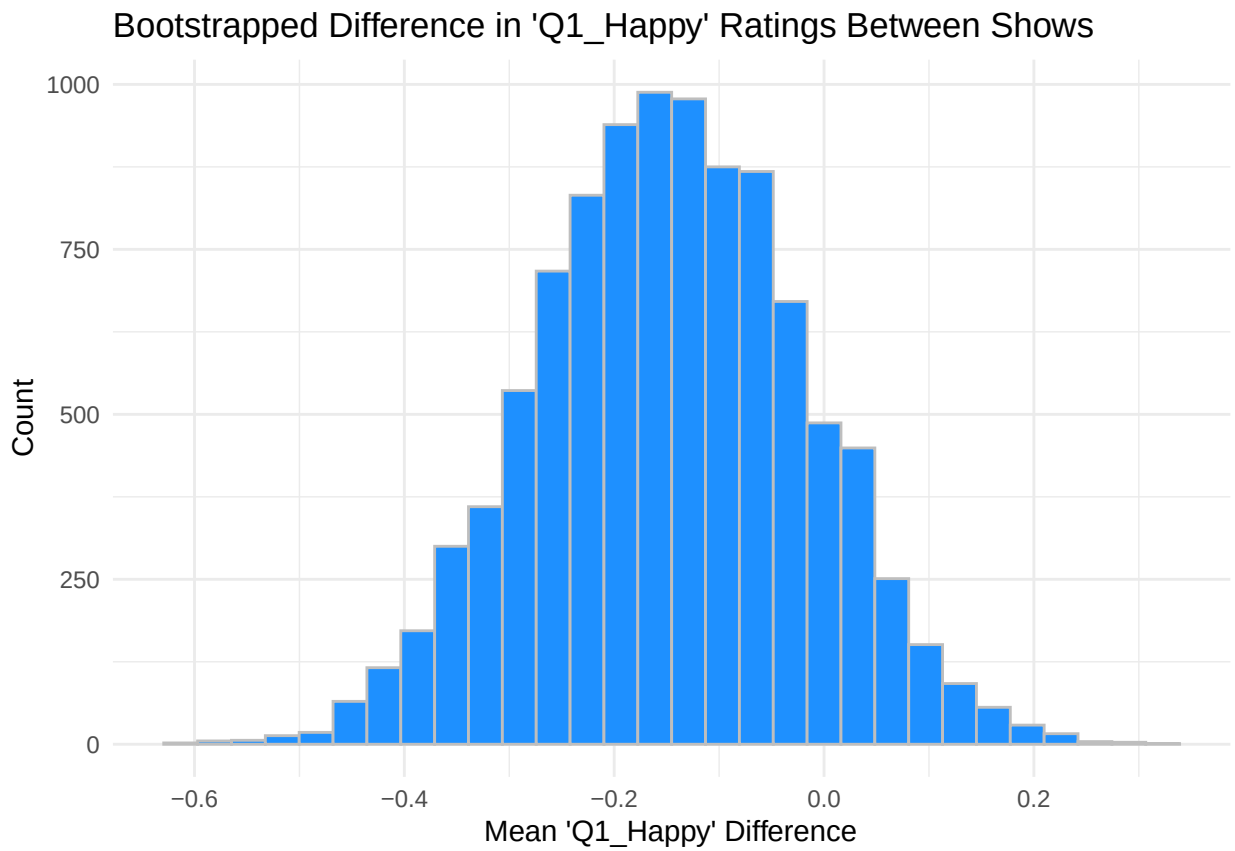



```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4164071 0.4527518  0.95 percentile 0.4288681
```

With 95% confidence, the proportion of all 2014 S-Class 550s that were painted black ranged from **0.417 to 0.453** when this data was collected. In other words, 41.7% to 45.3% of 2014 S-Class 550s listed in the market (from this data) were painted black.

Question 3 - NBC Pilot Surveys

Part A - Difference of Means in Happiness




```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.3979735 0.1001826  0.95 percentile -0.2424812
```

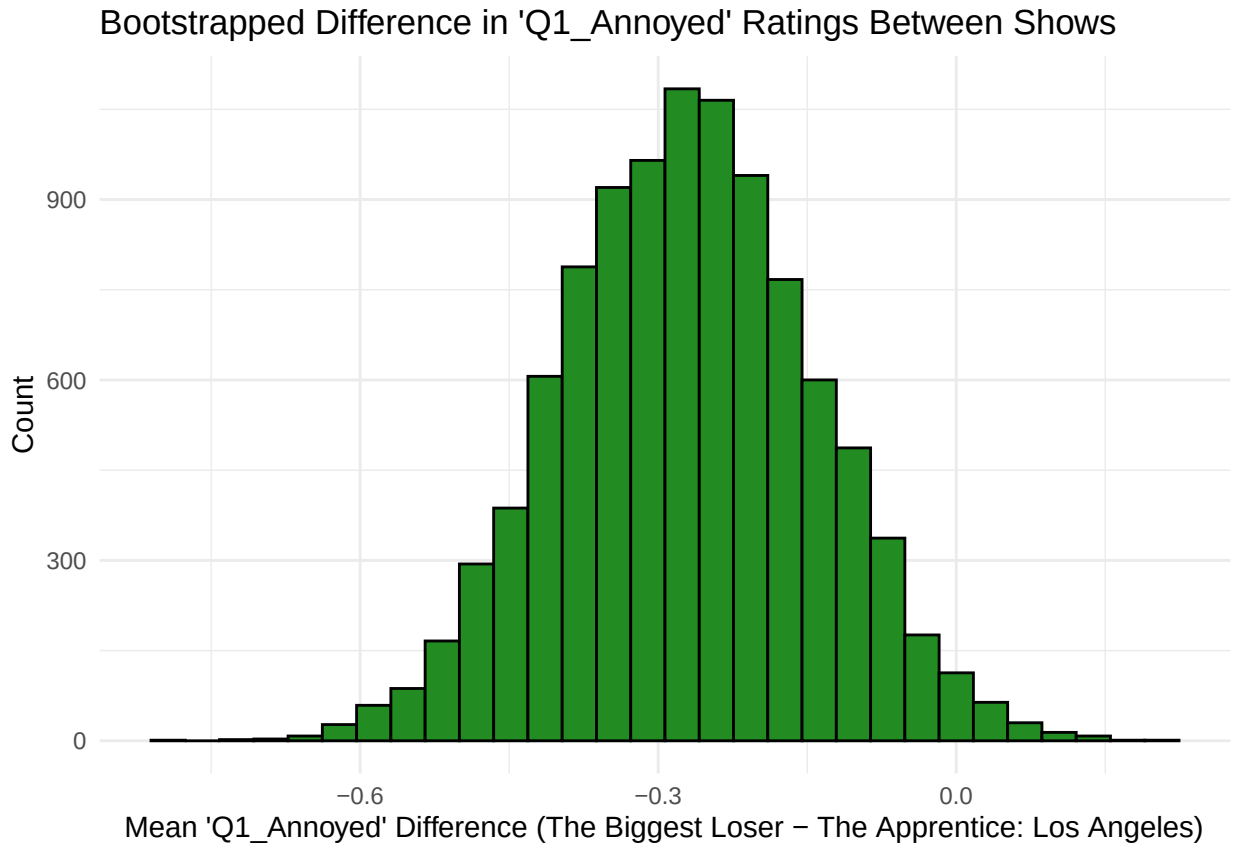
Question: Is there evidence that either (tv show names) "Living with Ed" or "My Name is Earl" consistently produced a higher mean "Q1_Happy" response among viewers?

Approach: First I created a subset of the original data that only contained viewer reponses from people who viewed "Living with Ed" or "My Name is Earl". Then, I resampled the difference of means between the "Q1_Happy" response of both shows 10000 times. I also created a ggplot to visualize the bootstrapped mean differences in the "Q1_Happy" ratings between the two shows. Using that bootstrapped data, I constructed a 95% confidence interval to list the range of mean difference between the TV shows' "Q1_Happy" response over the 1000 bootstrap samples.

Results: With 95% confidence, the mean difference in responses of "Q1_Happy" for viewers who watched "Living with Ed" and viewers who watched "My Name is Earl" ranged from -0.400 to 0.102. Because a possible difference of zero is contained within this confidence interval, there is no statistical significance (at the 5% significance level) to claim that either show consistely produced a higher mean happiness response among viewers.

Conclusion: This question is **unsupported** by the 95% confidence interval, as there is no data to support a statistically significant mean difference between "Q1_Happy" response of the tv shows' "Living with Ed" and "My Name is Earl".

Part B - Difference of Means in Annoyingness




```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5176008 -0.02264907  0.95 percentile -0.270997
```

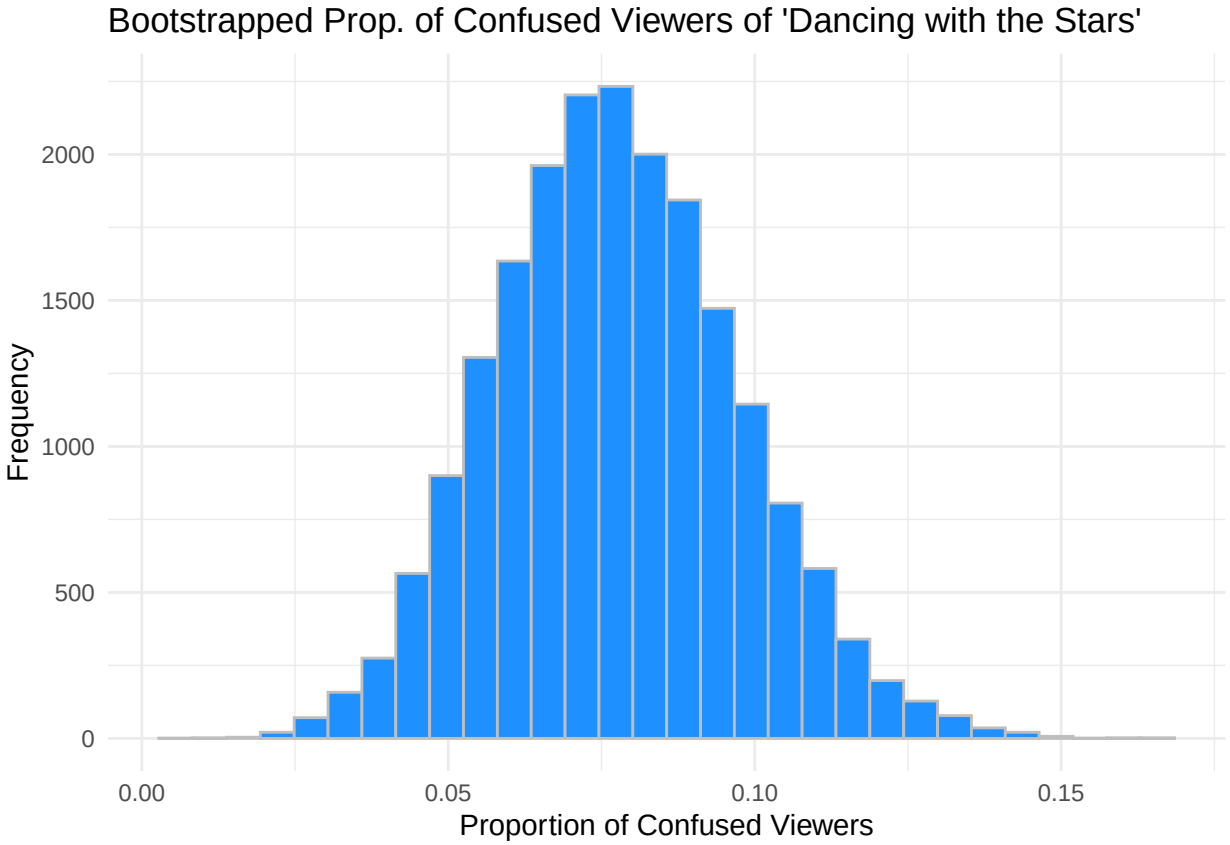
Question: Is there evidence that either (tv show names) "The Biggest Loser" or "The Apprentice: Los Angeles" consistently produced a higher mean "Q1_Annoyed" response among viewers?

Approach: First I created a subset of the original data that only contained viewer reponses from people who viewed "The Biggest Loser" or "The Apprentice: Los Angeles". Then, I resampled the difference of means between the "Q1_Annoyed" response of both shows 10000 times. I also created a ggplot to visualize the bootstrapped mean differences in the "Q1_Annoyed" ratings between the two shows. Using that bootstrapped data, I constructed a 95% confidence interval to list the range of mean difference between the TV shows' "Q1_Annoyed" response over the 1000 bootstrap samples.

Results: With 95% confidence, the mean difference in responses of "Q1_Annoyed" for viewers who watched "The Biggest Lose" and viewers who watched "The Apprentice: Los Angeles" ranged from -0.523 to -0.016. Because zero is not contained within the 95% confidence interval, there is statistical significance (at the 5% significance level) to claim that either "The Biggest Loser" or "The Apprentice: Los Angeles" consistently produced a higher mean "Q1_Annoyed" response among viewers.

Conclusion: The question is **supported** by the 95% confidence interval, as there is data to support a statistically significant mean difference (at the 5% level) between "Q1_Annoyed" response of the tv shows' "The Biggest Loser" and "The Apprentice: Los Angeles". Specifically, because the 95% confidence interval range is in the negatives, the evidence supports that "The Apprentice: Los Angeles" made viewers more annoyed compared to viewers of "The Biggest Loser".

Part C - Proportion of Confusion in Dancing with the Stars




```
##        name      lower     upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.1160221  0.95 percentile 0.09392265
```

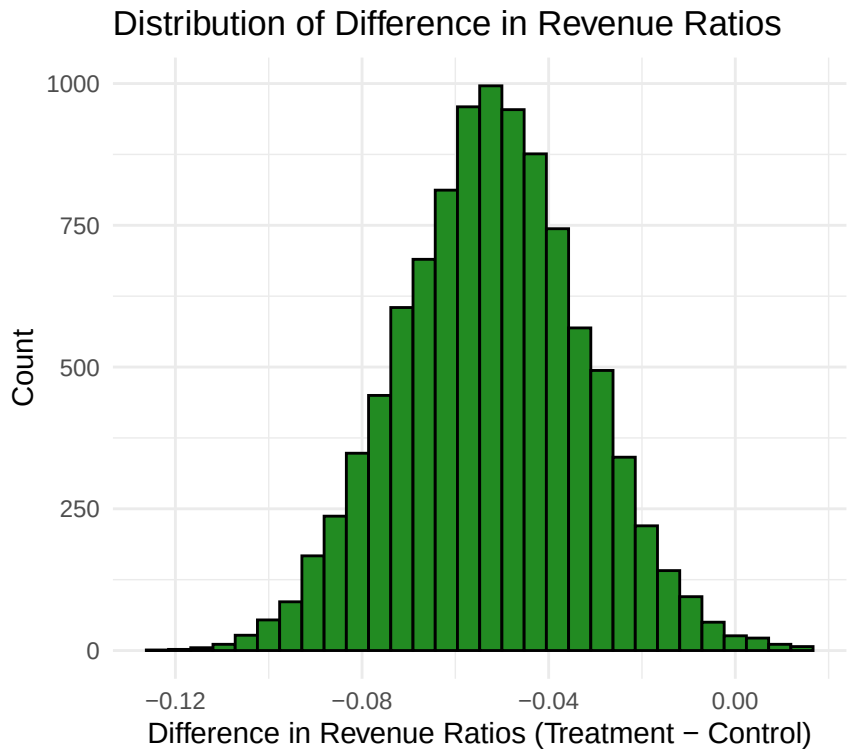
Question: What proportion of American TV watchers would we expect to give a response of 4 or greater to the 'Q2_Confusing' question, where 4 indicates agreeing and 5 indicates strongly agreeing with the statement that 'Dancing with the Stars' is a confusing show?

Approach: First, I filtered the data set to only include reponses for the show "Dancing with the Stars". Then, I used the mutate() function to create a new variable, isConfused, which marks respondents who gave a rating of 4 or 5 on Q2_Confusing as "Y" (confused) and those who gave a rating of 1, 2, or 3 as "N" (not confused). Afterwards, I resampled the proportion of respondents who found the show confusing (those with "Y" in isConfused) across 20,000 bootstrap samples using the prop() function. This allows for estimating the proportion of viewers who would give a response of 4 or greater to the "Q2_Confusing" question. I also created a ggplot to visualize the bootstrap distribution of the proportions of the confused viewers (with the response of 4 or greater) for the show 'Dancing with the Stars'. Lastly, I constructed the 95% confidence interval to list the proportion of viewers who gave a response of 4 or greater to the "Q2_Confusing" Question over the 1000 bootstrap samples.

Results: With 95% confidence, the proportion of viewers who gave a response of 4 or greater to the "Q2_Confusing" question is between 0.039 and 0.116. Because zero is not contained within the 95% confidence interval, there is statistical significance (at the 5% significance level) here.

Conclusion: The question is **supported** by the 95% confidence interval, as there is data to support a statistically significant proportion of viewers who gave a response of 4 or greater to the "Q2_Confusing" question for the show, "Dancing with the Stars", meaning that there was some sort of confusion for this show.

Problem 4 - EBay




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.09052404 -0.01322533  0.95 percentile -0.05228145
```

Question: Is there a difference in revenue ratio between DMAs with paid search advertising and DMAs without paid search advertising, or more specifically, is there evidence to support that DMAs with paid search advertising bring in more revenue than DMAs without paid search advertising?

Approach: First, I created a new variable, rev_ratio, by dividing the revenue after the experiment (rev_after) by the revenue before the experiment (rev_before) for each designated market area (DMA) in the dataset. This ratio captures the change in EBay's revenue for each DMA during the experimental period.Next, I performed a bootstrap simulation using 10,000 resamples to compute the difference in mean revenue ratios between the treatment group (where ads were paused) and the control group (where ads continued). I also created a ggplot to visualize the bootstrap distribution of the mean difference in revenue ratios between the treatment and control groups. Finally, I constructed a 95% confidence interval for the difference in revenue ratios.

Results: With 95% confidence, the difference in means of the revenue ratio between DMAs with paid search advertising and DMAs without paid search advertising ranged from -0.089 and -0.013 dollars. Because zero (dollars) isn't contained in this confidence interval, there is evidence (at the 5% significance level) to support a difference in revenue ratio between DMAs with paid search advertising and DMAs without paid search advertising.

Conclusion: The question is **supported** by the 95% confidence interval, as there is data to support a statistically significant mean difference between the revenue ratio of DMAs with paid search advertising and DMAs without paid search advertising. Specifically, since the confidence interval contained only negative amounts of dollars, this means that DMAs with paid search advertising had a higher revenue ratio than DMAs without paid search advertising (as the difference of means was treatment - control, so the revenue ratio was lower in the treatment-group DMAs because the difference of means was negative).