Homework 3

Anthony Zhang - SDS 315 UT Austin, az22589, https://github.com/antzha630/HW-3

Contents

Github Link for R Script

---

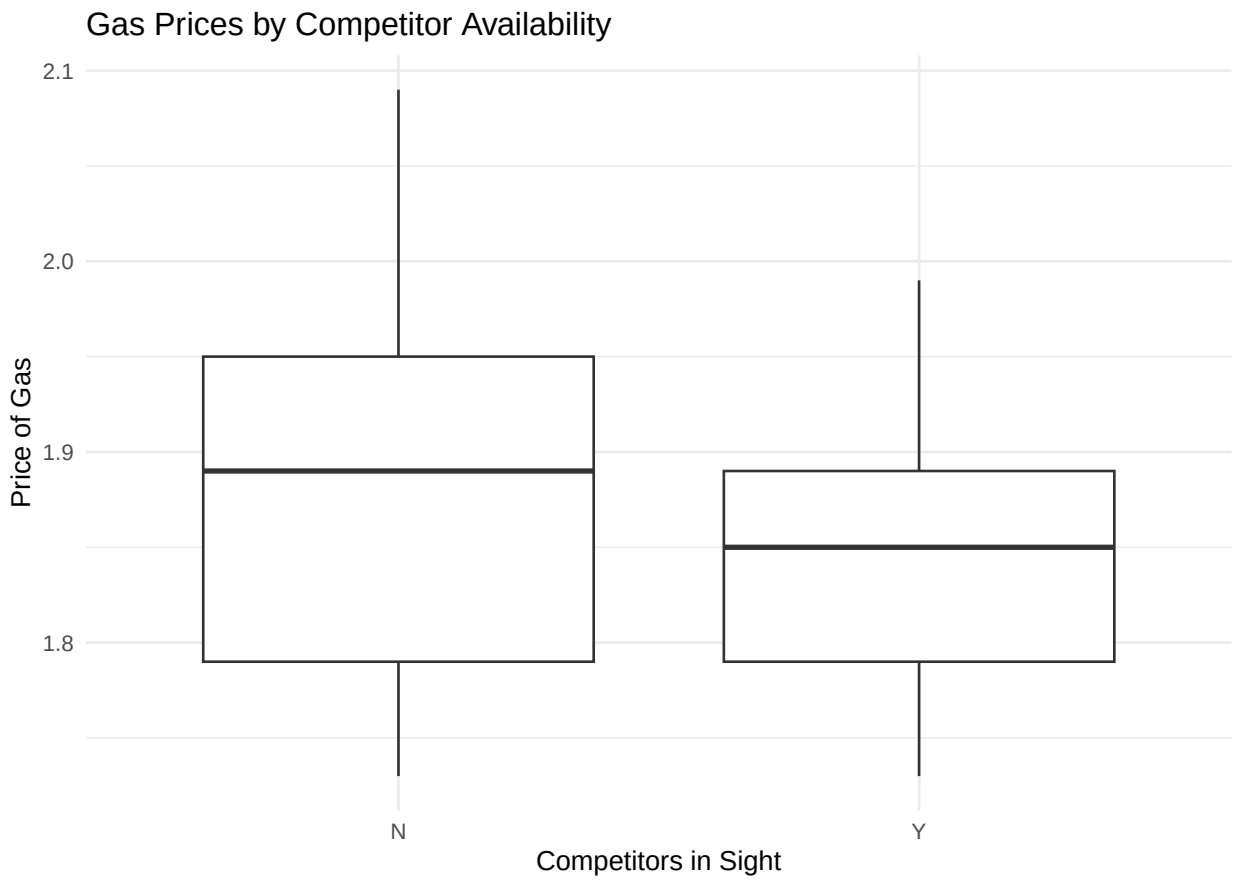
1. Gas Prices Analysis

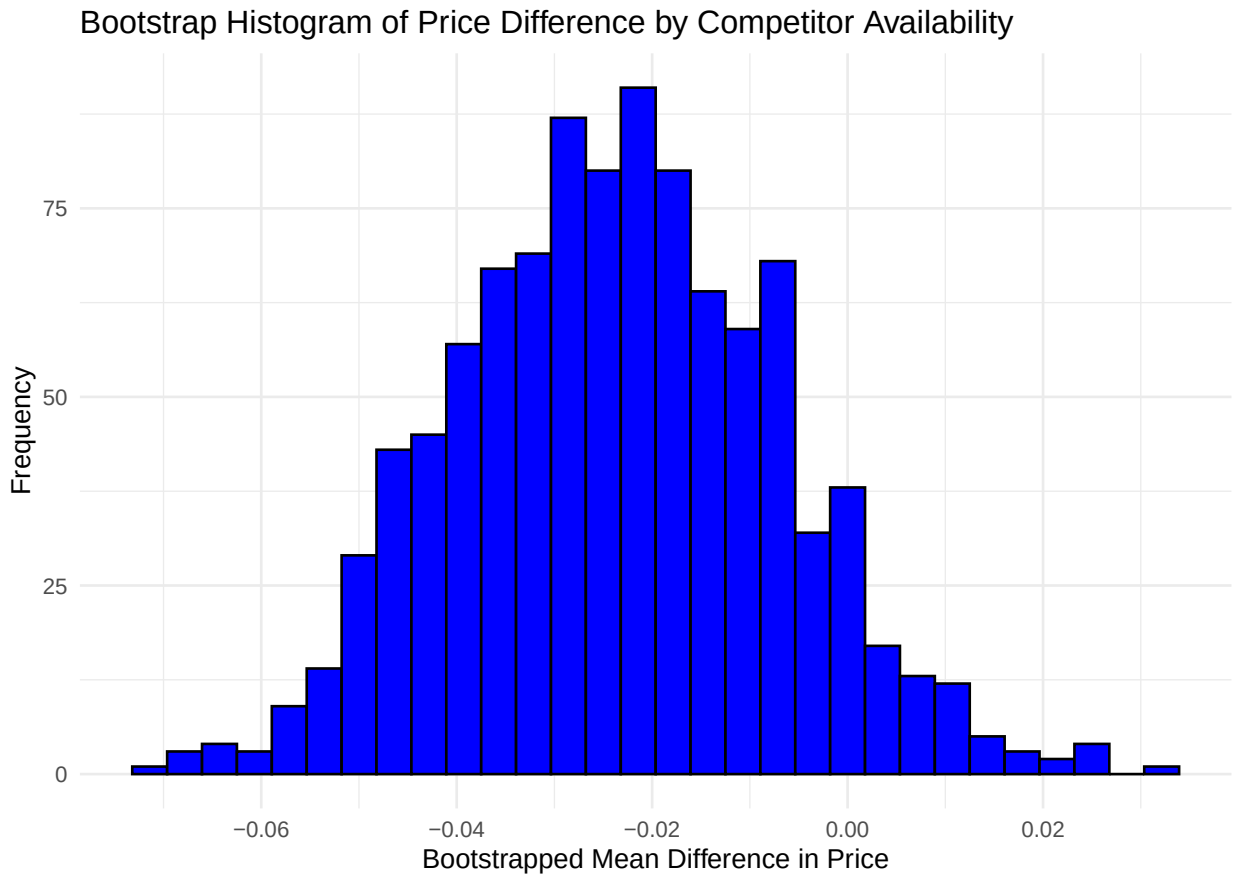
Theory A: Gas stations charge more if they lack direct competition in sight.

Claim:

Gas stations with no direct competition charge higher prices compared to gas stations with direct competition nearby.

Evidence:


Gas Prices by Competitor Availability
2.1
2.0
1.9
1.8
Price of Gas
N
Y
Competitors in Sight




```
##   name       lower       upper level     method    estimate
## 1    Y -0.05400336 0.009152595  0.95 percentile -0.02348235
```

This boxplot shows the price of gas by whether there are competitors nearby. Gas stations with no competitors nearby have a higher median gas price compared to gas stations with competitors nearby by about 0.05 dollars. However, gas stations with no competitors nearby have a higher IQR than gas stations with competitors nearby by about 0.06. As a result, there is more variability in their gas prices.

My histogram of bootstrapped mean price differences by competitor availability is shown above. More of the bootstrapped mean differences in price seem to be grouping around -0.025 while there are less bootstrapped mean differences in price around the sides.

I used bootstrapping to estimate the difference in means for gas prices between gas stations with or without direct competitors nearby. I resampled the dataset 1000 times and calculated the difference in means for each resample with replacement. Then, I constructed a 95% confidence interval based on the bootstrap histogram distribution. The difference in mean gas price between gas stations with competitors nearby and no competitors nearby is between -0.056 and 0.008, with 95% confidence. Since the 95% confidence interval contains 0, we cannot confidently claim there is a difference in gas prices between whether there is competitors nearby.

Conclusion:

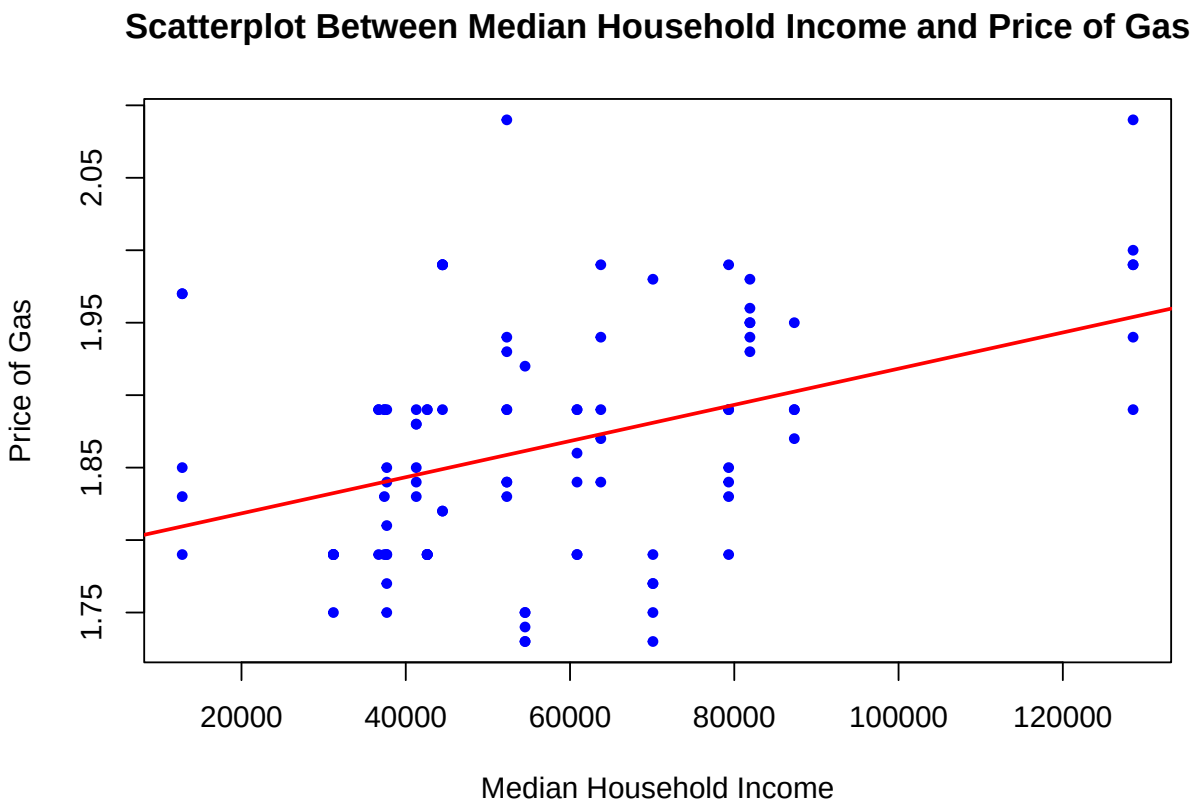
Because the confidence interval contains zero, evidence doesn't strongly support the theory that gas stations charge more if they lack direct competition in sight.

Theory B: The richer the area, the higher the gas prices.

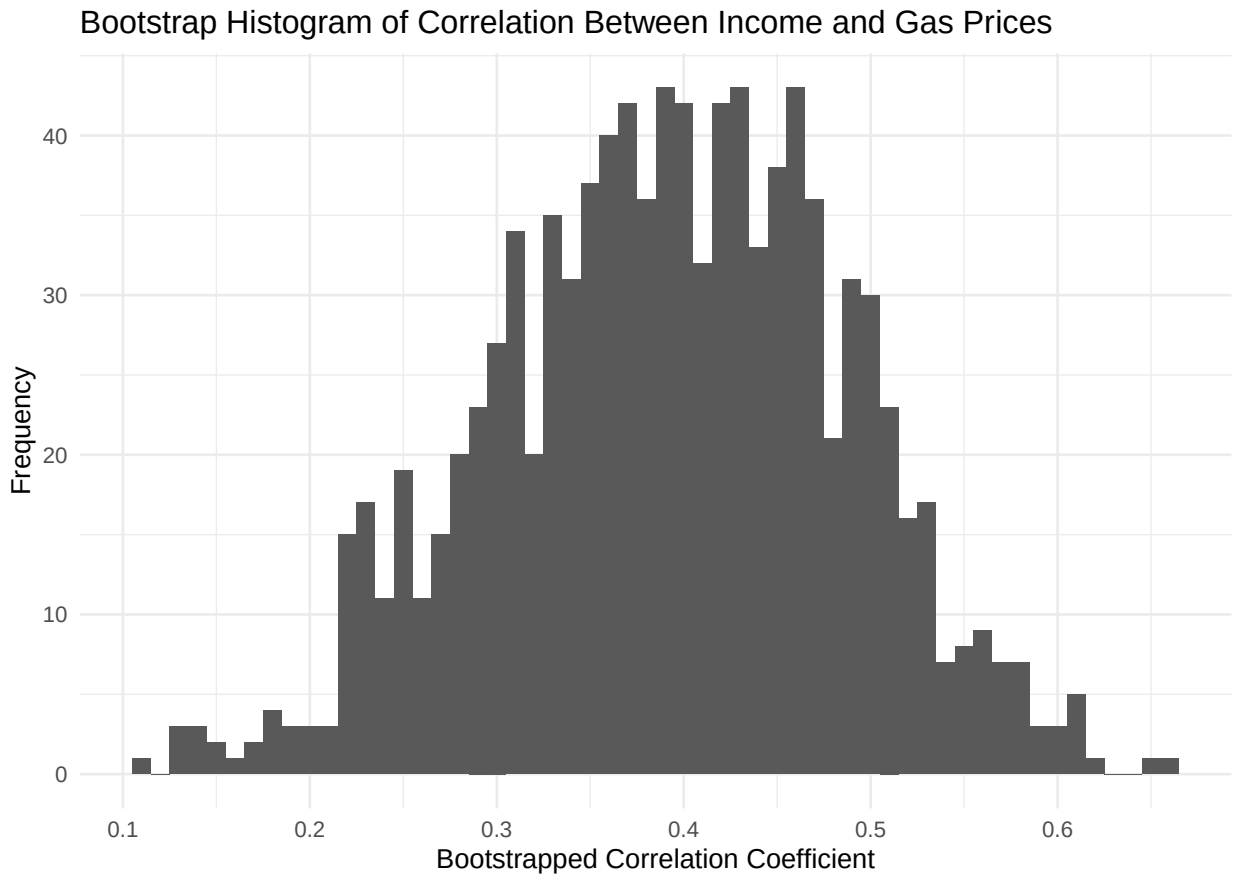
Claim:

Gas stations that are in richer areas charge higher gas prices.

Evidence:




```
## [1] 0.4
```




```
##   name     lower    upper level     method  estimate
## 1  cor 0.2153018 0.568798  0.95 percentile 0.3961546
```

This is a scatterplot between median household income of the ZIP code and the price of gas. There is a correlation coefficient of 0.4 between the two variables. Because the points seem decently scattered and the correlation coefficient is pretty close to 0.5, we are say there is a moderate positive correlation between these two factors. A positive correlation indicates that a higher median household income leads to a higher price of gas.

My histogram of bootstrapped correlation coefficient between income and price of gas is shown above. More of the bootstrapped correlation coefficients seem to be grouping around 0.45 while there are less bootstrapped mean differences in price around the sides.

I used bootstrapping to estimate the correlation between income and gas price. With 95% confidence, the true correlation between median household income and price of gas is between 0.209 and 0.579. Since this confidence interval doesn't contain 0, we can say there is a significant positive relationship between income and gas price. There is a small to moderate effect size because of the correlation coefficient in the range of 0 to 0.5 which means that income does have an effect on price of gas but it isn't a big one.

Conclusion:

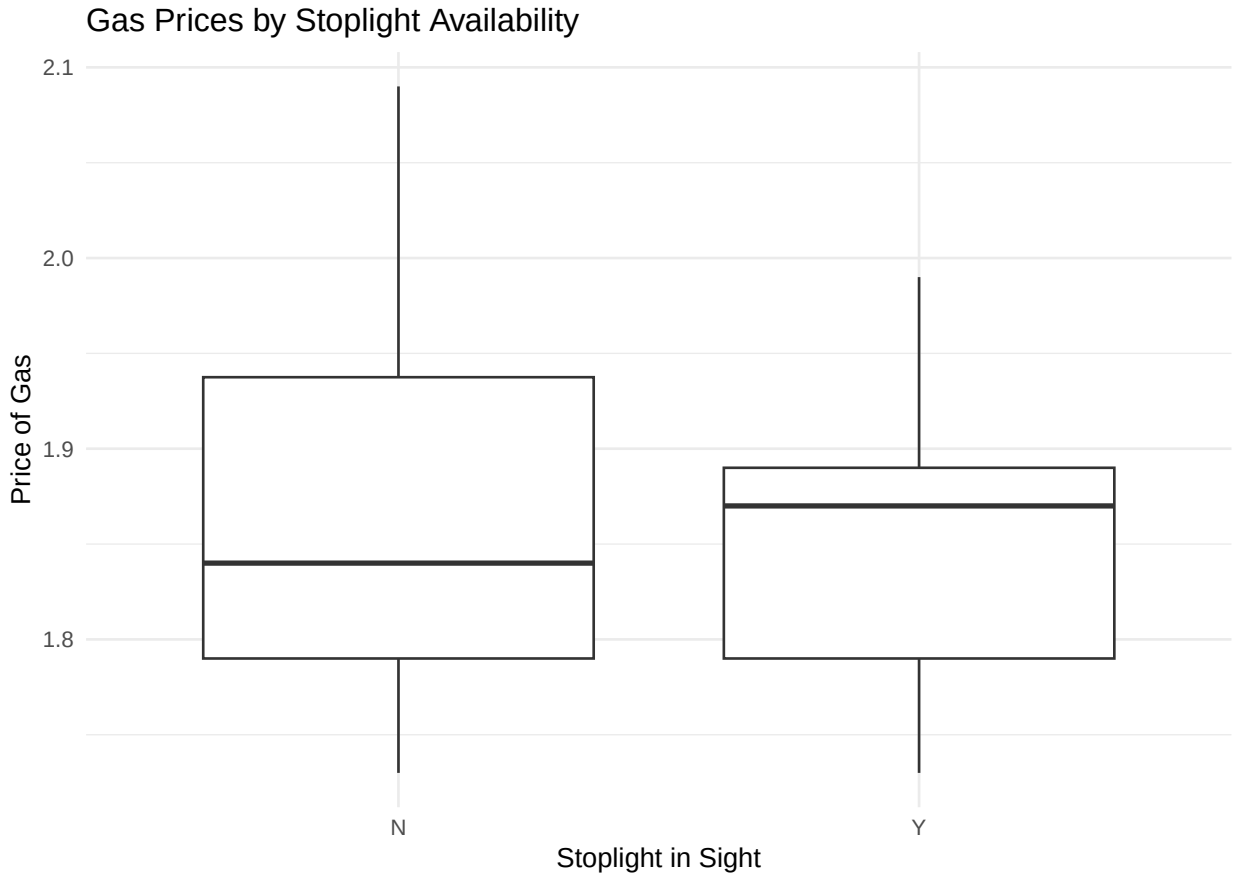
Since the confidence interval is entirely positive, evidence supports the theory that a higher median household income leads to a higher price of gas. This relationship is at a small to moderate effect size and there is a statistically significant relationship.

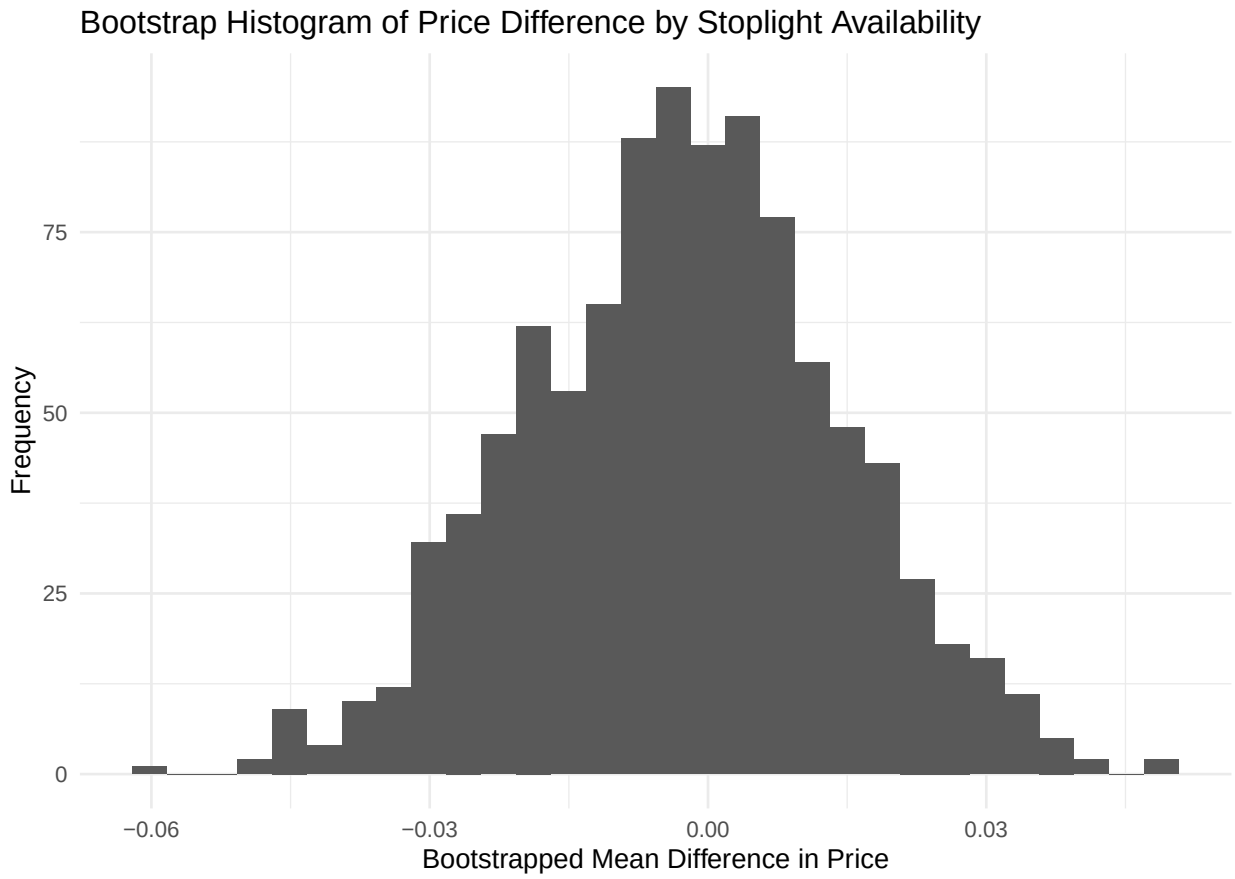
Theory C: Gas stations at stoplights charge more.

Claim:

Gas stations at stoplights charge more than gas stations that aren't at stoplights.

Evidence:






```
##   name       lower      upper level     method     estimate
## 1    Y -0.03583844 0.03041393  0.95 percentile -0.003299916
```

This boxplot shows the price of gas by whether there are stoplights nearby. Gas stations with stoplights nearby have a higher median gas price compared to gas stations with no stoplights nearby by about 0.04 dollars. However, gas stations with no stoplights nearby have a higher IQR than gas stations with stoplights nearby by about 0.05 dollars. As a result, there is more variability in their gas prices.

My histogram of bootstrapped mean price differences by stoplight availability is shown above. More of the bootstrapped mean differences in price seem to be grouping around -0.01 while there are less bootstrapped mean differences in price around the sides.

I used bootstrapping to estimate the difference in means for gas prices between gas stations with or without direct stoplights nearby. The difference in mean gas price between gas stations with stoplights nearby and no stoplights nearby is between -0.038 and 0.028, with 95% confidence. Since the 95% confidence interval contains 0, we cannot confidently claim there is a difference in gas prices depending on whether there is stoplights nearby.

Conclusion:

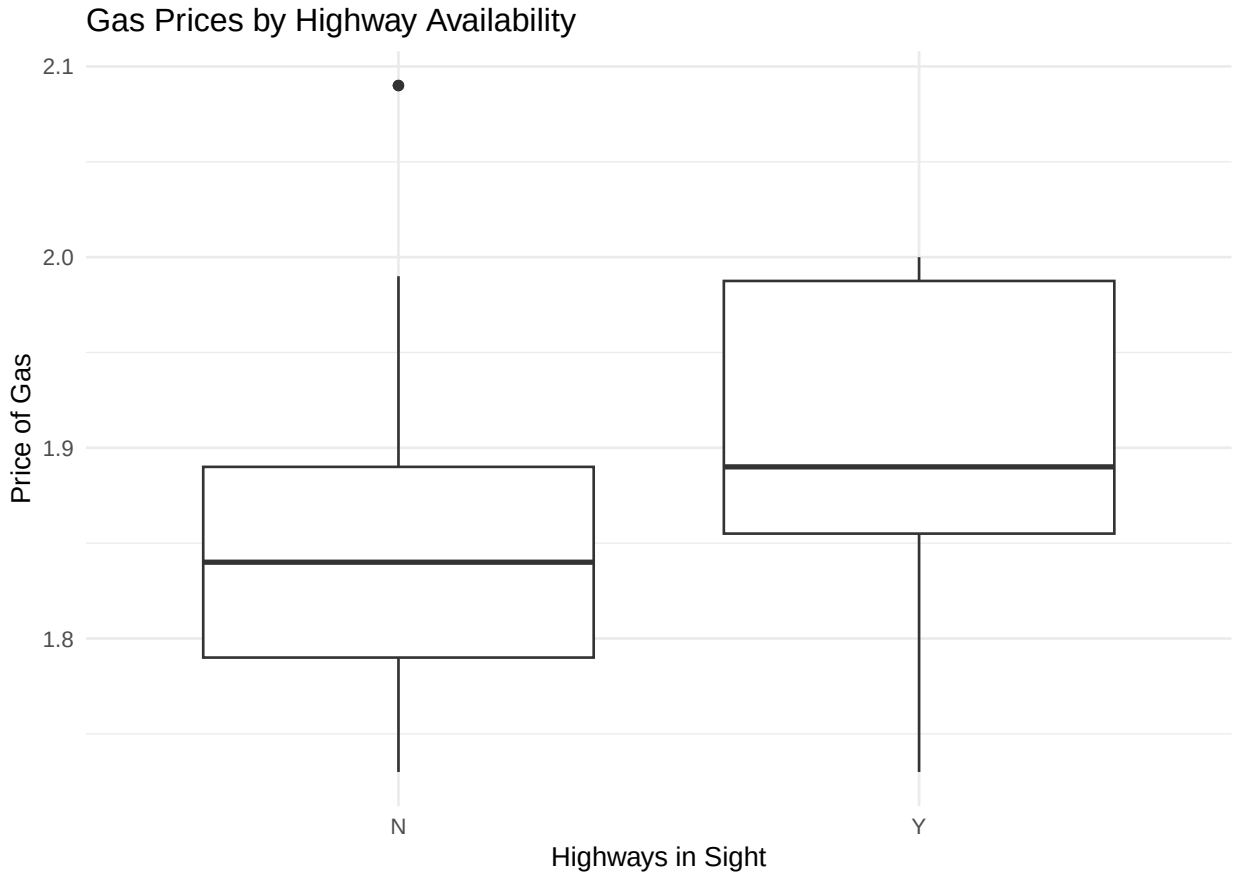
Because the confidence interval contains zero, evidence doesn't strongly support the theory that gas stations charge more at gas stations with stoplights compared to gas stations without stoplights.

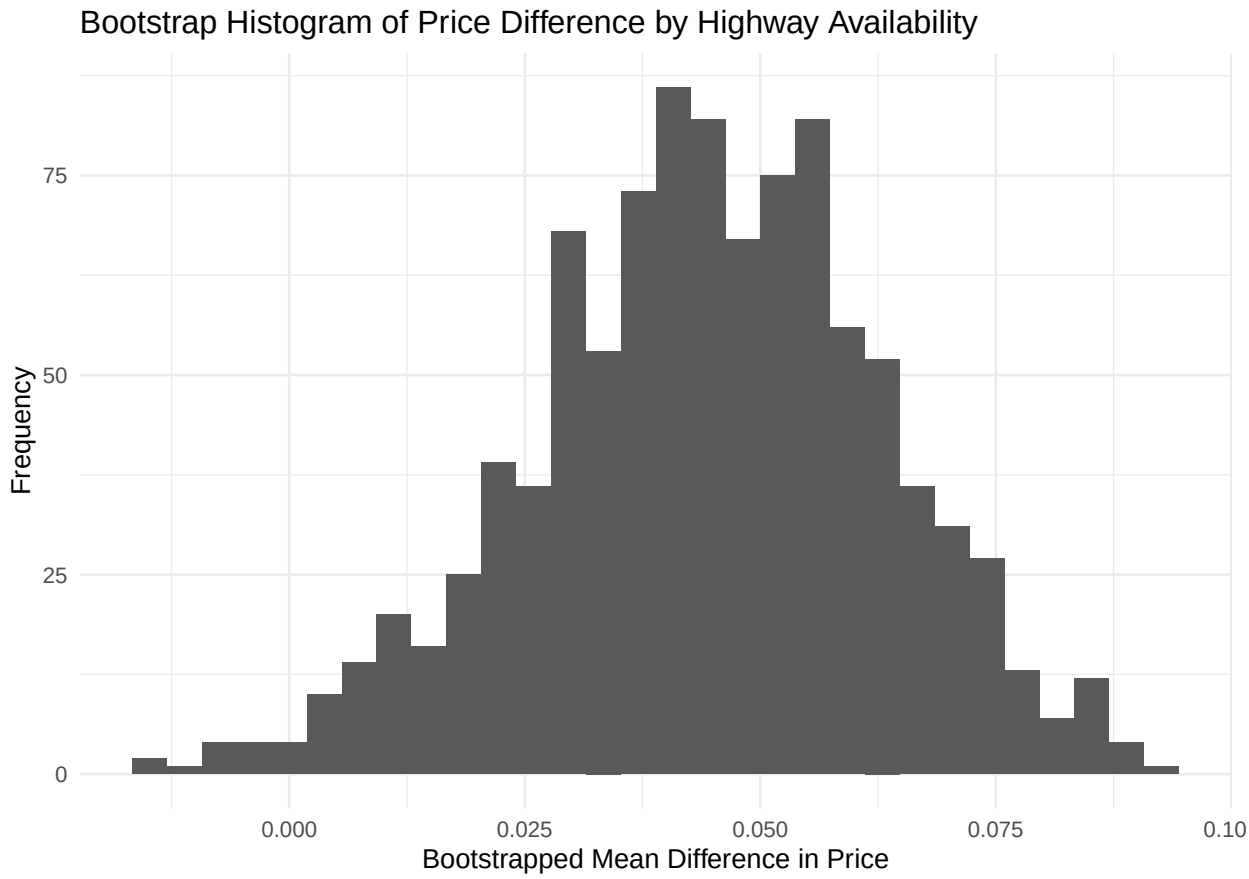
Theory D: Gas stations with direct highway access charge more

Claim:

Gas stations with direct highway access charge more for gas compared to gas stations without direct highway access.

Evidence:






```
##   name       lower      upper level     method  estimate
## 1    Y 0.006004101 0.07897339  0.95 percentile 0.0456962
```

This boxplot shows the price of gas by whether the gas station has direct access to highways or not. Gas stations with direct access to highways have a higher median gas price compared to gas stations with no direct access to highways by about 0.05 dollars. Gas stations with direct access to highways have a higher IQR than gas stations with no direct access to highways by about 0.04 dollars. As a result, there is more variability in their gas prices.

My histogram of bootstrapped mean price differences by direct highway access availability is shown above. More of the bootstrapped mean differences in price seem to be grouping around 0.05 while there are less bootstrapped mean differences in price around the sides.

I used bootstrapping to estimate the difference in means for gas prices between gas stations with or without direct access to highways. The difference in mean gas price between gas stations with direct access to highways or no direct access to highways nearby is between 0.009 and 0.084, with 95% confidence. Since the 95% confidence interval doesn't contain 0, we can confidently claim there is a difference in gas prices depending on whether it has direct access to highways or not. There is a small to moderate effect size because of the size of the confidental interval.

Conclusion:

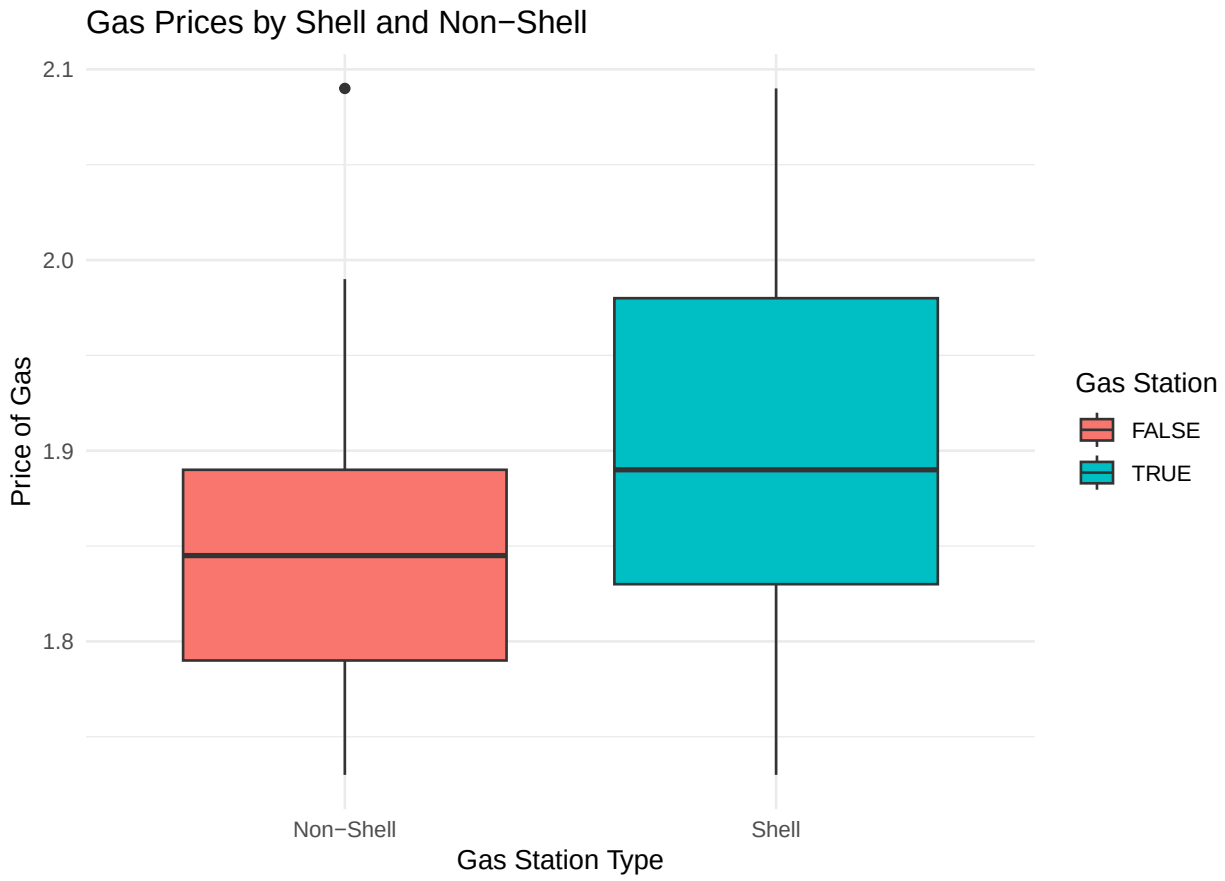
Since the confidence interval is entirely positive, evidence supports the theory that gas stations with direct highway access charge more. This relationship is a small to moderate effect size and there is a statistically significant relationship.

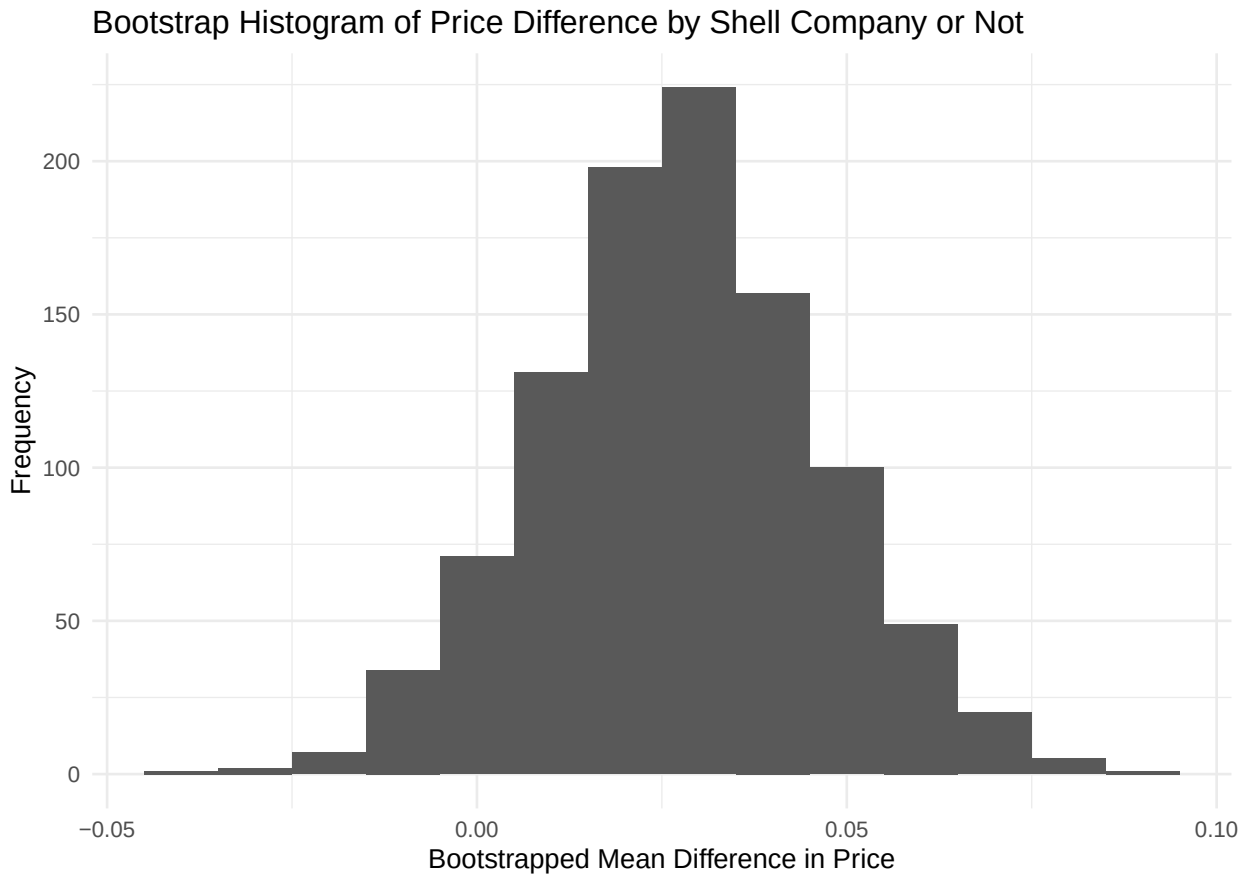
Theory E: Shell charges more than all other non-Shell brands.

Claim:

Shell has higher gas prices compared to all the other non-Shell brands.

Evidence:






```
##     name        lower      upper level     method   estimate
## 1 result -0.009470846 0.06506687  0.95 percentile 0.02740421
```

This boxplot shows the price of gas by whether the gas station company is Shell or not Shell. Gas stations that are Shell have a higher median gas price compared to gas stations that are not Shell by about 0.05 dollars. Also, gas stations that are Shell have a higher IQR than gas stations that are not Shell by about 0.05 dollars. As a result, there is more variability in their gas prices.

My histogram of bootstrapped mean price differences by whether it is Shell or not is shown above. More of the bootstrapped mean differences in price seem to be grouping around 0.003 while there are less bootstrapped mean differences in price around the sides.

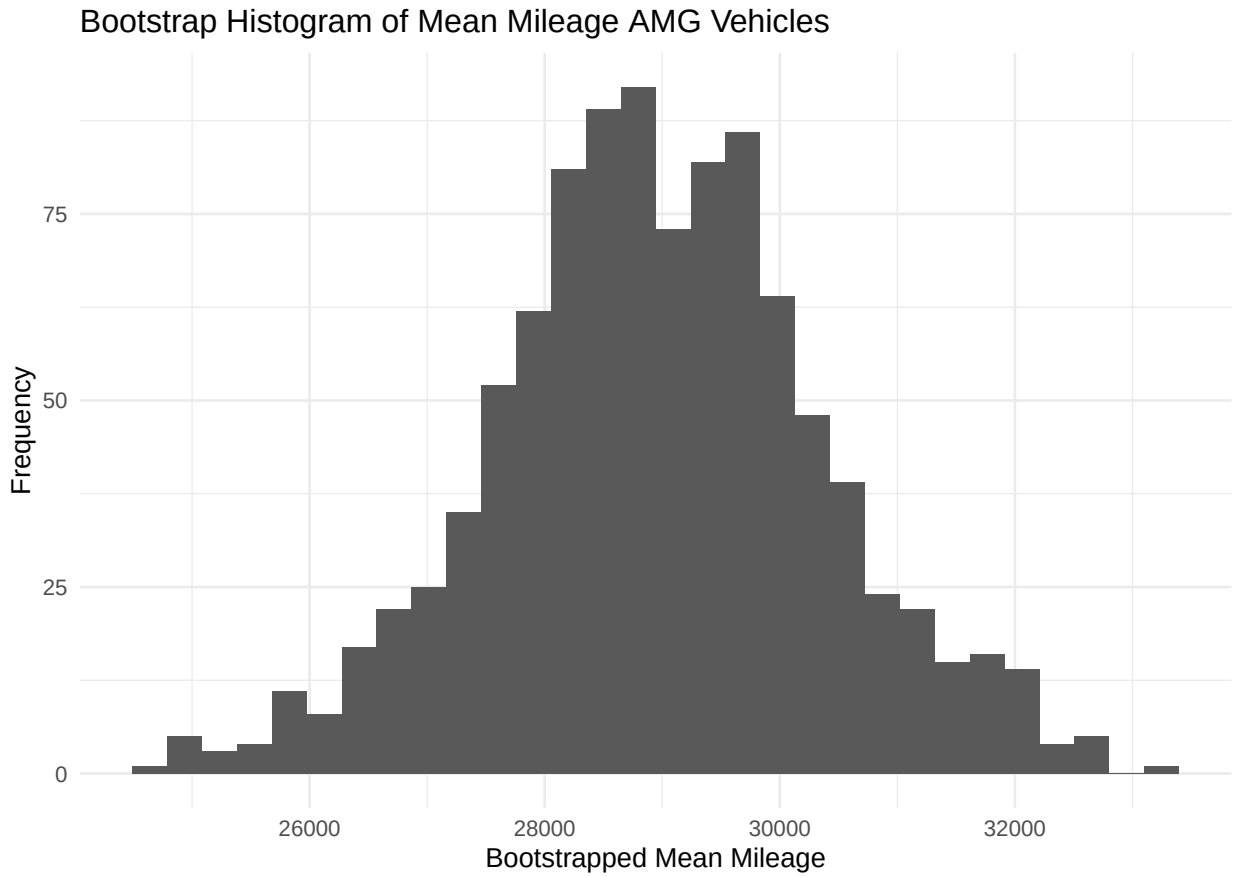
I used bootstrapping to estimate the difference in means for gas prices between gas stations that are Shell and gas stations that aren't Shell. The difference in mean gas price between these two is between -0.009 and 0.068, with 95% confidence. Since the 95% confidence interval contains 0, we cannot confidently claim there is a difference in gas prices depending on whether it is Shell or not.

Conclusion:

Because the confidence interval contains zero, evidence doesn't strongly support the theory that Shell gas stations charge more than non Shell gas stations.

2. Mercedes S Class Cars

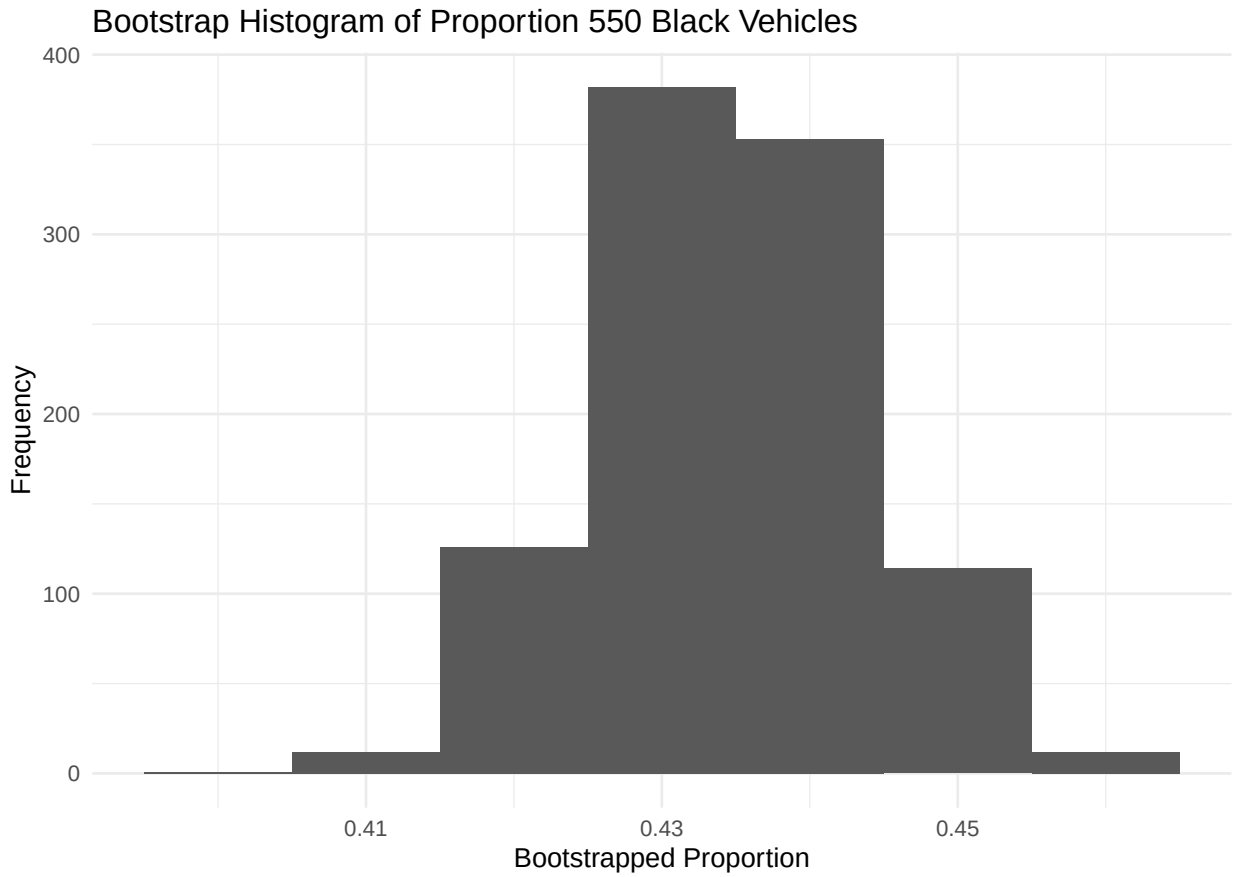
PART A




```
##   name    lower    upper level     method estimate
## 1 mean 26035.44 31863.46  0.95 percentile 28997.34
```

With 95% confidence, we estimate that the true average mileage of 2011 S-Class 63 AMG vehicles hitting the used-car market falls between 26367 miles and 31754 miles. If a group of people or an individual person is looking to buy this type of car at the used car market, they can expect it to be between this rank. It is important to know that it won't 100% be in this range and we expect it to be wrong 5% of the time.

PART B




```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4174455 0.4527605  0.95 percentile 0.4347525
```

With 95% confidence, we estimate that the true proportion of 2014 S-Class 550 vehicles that were painted black falls between 0.42 and 0.45. This shows that a big amount of luxury cars on the used market were black. This makes sense because industry trends show this is a common color for expensive cars since it can be used by organizations like the FBI, government, criminals, or people in general. It is a slick and clean color that matches with many different enviornments and places.

3. TV Network

PART A

Question:

Does “Living with Ed” or “My Name is Earl” result in higher happiness ratings among viewers?

Approach:

First, I filtered the dataset to only have responses from the two shows that are listed above. I used bootstrapping to construct a 95% confidence interval for the difference in mean Q1_Happy ratings between the two shows.

```
##         name      lower     upper level     method   estimate
## 1 diff_means -0.3960883 0.1268597  0.95 percentile -0.1490515
```

Results: The 95% confidence interval for the difference in mean happiness ratings (Living with Ed - My Name is Earl) is (-0.41, 0.07). Since the interval contains 0, this doesn't show a significant difference in happiness ratings between the two shows.

Conclusion: Since the confidence interval contains 0, there is no evidence supporting the idea that one show is consistently more happy than the other.

PART B

Question: Does "The Biggest Loser" or "The Apprentice: Los Angeles" result in higher annoyance ratings among viewers?

Approach:

First, I filtered the dataset to only have responses from the two shows that are listed above. I used bootstrapping to construct a 95% confidence interval for the difference in mean Q1_annoying ratings between the two shows.

```
##         name      lower       upper level     method  estimate
## 1 diff_means -0.5053846 -0.02910906  0.95 percentile -0.270997
```

Results: The 95% confidence interval for the difference in mean annoyance ratings (Biggest Loser - Apprentice: LA) is (-0.52, -0.03). Since the interval doesn't contain 0, this shows a significant difference in annnoyance ratings between the two shows.

Conclusion: Since the confidence interval is entirely negative, there is evidence supporting the idea that "The Apprentice: Los Angeles" is consistently more annoying to viewers than "The Biggest Loser".

PART C

Question: What proportion of viewers would rate the show "Dancing with the Stars" as confusing on the Q2_Confusing question?

Approach:

First, I filtered the dataset to only have responses from the show that is listed above. I used bootstrapping to construct a 95% confidence interval for the proportion of respondents who rated Q1_confusing a 4 or more (out of 5).

```
##        name      lower    upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.121547  0.95 percentile 0.07734807
```

Results: The 95% confidence interval for the proportion of respondents who rated 4 or more is (0.04, 0.12).

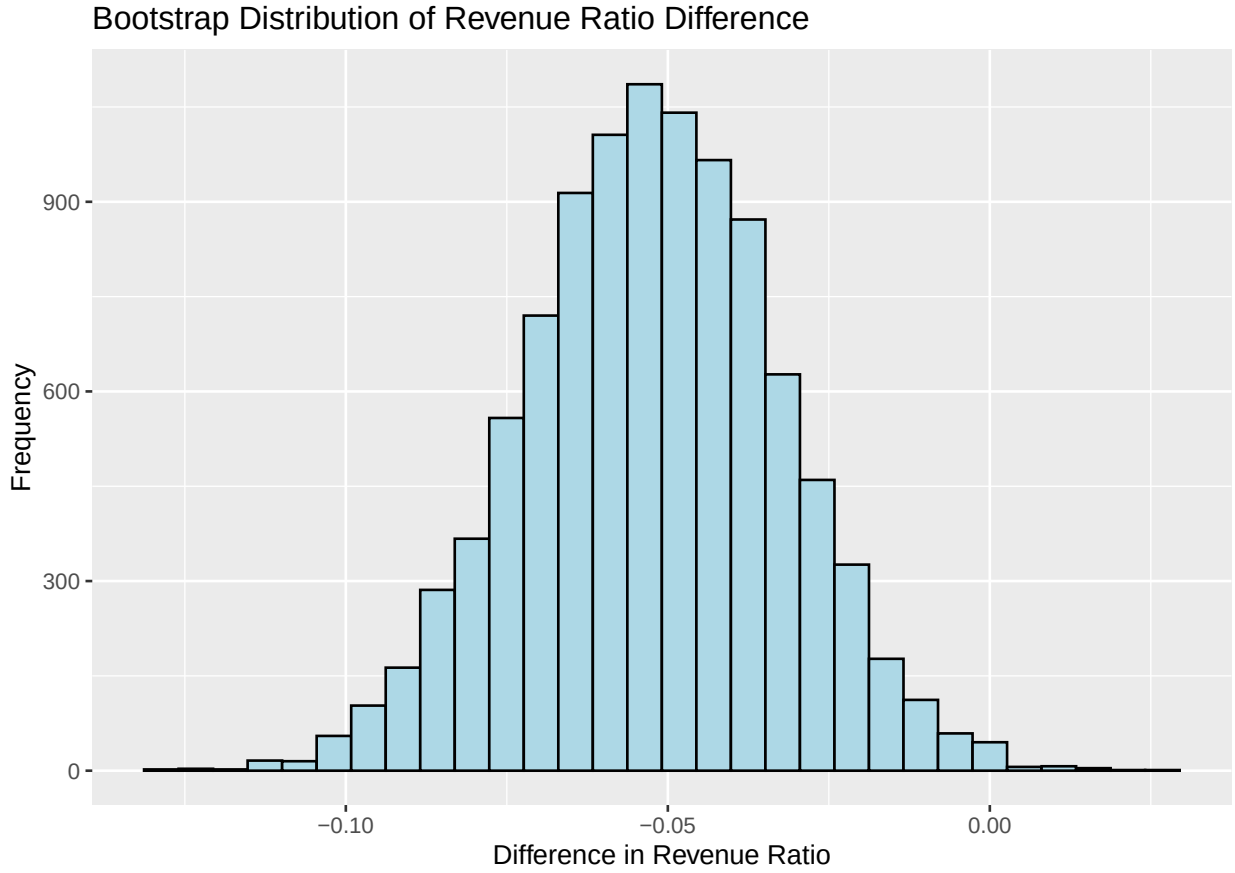
Conclusion: Because the proportion of viewers who confused the show confusing is on the lower side (less than 0.15), there isn't a big portion of the audience that struggles to understand the show and it isn't that confusing.

4. Ebay

Question:

Does the extra traffic brought to Ebay from paid search results—above and beyond what we'd see if we "went organic"— justify the cost of the ads themselves? Is the revenue ratio for the treatment group or the group that has organic search results less than the revenue ratio for the control group or the paid search results?

Approach: First, I made a new column called revenue ratio which is just the revenue after divided by the revenue before. I did this for each DMA which is the Designated Market Area. Then, I used the bootstrap method with 10000 samples with replacement to find the difference in mean revenue ratios between the treatment group and control group. Both of these are stated in the problem and above. This was used to make a 95% confidence interval for the difference in mean revenue ratios.




```
##         name       lower       upper level     method    estimate
## 1 diff_means -0.09157858 -0.01407298  0.95 percentile -0.05228145
```

Results: The 95% bootstrap confidence interval for the difference in mean revenue ratios was (-0.09, -0.01). Since this confidence interval doesn't include 0, there is a statistically significant difference in mean revenue ratios.

Conclusion: Because the confidence interval only has negative values, there is evidence supporting the fact that Ebay's paid advertising on Google search bars generated aditional revenue. The effect size is pretty small due to how small the interval is so it is up to Ebay to decide whether they want to take the hit and take a look at their options.