homework 3

anika shareef

2025-02-05

as236452

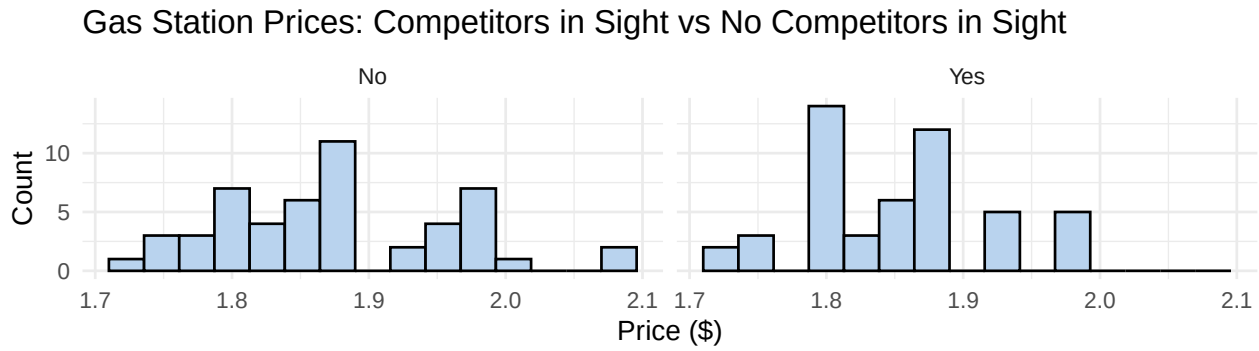
Github Link: https://github.com/anikashareef/sds315_hw3.git

Problem 1:

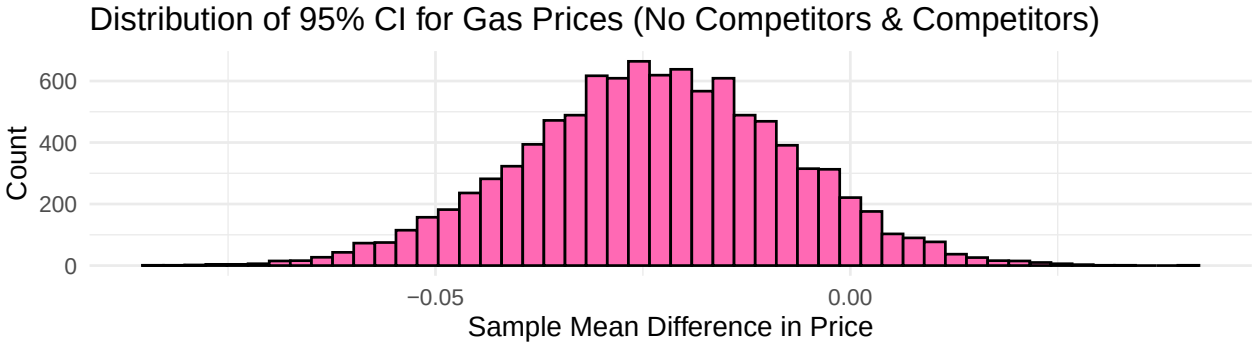
A)

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

Evidence:




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05521847 0.007595214  0.95 percentile -0.02348235
```



Before conducting the bootstrap, I made a graph to visualize the difference between gas prices between gas stations without competitors in sight and gas stations with competitors in sight. Just based on these graphs, it appears that stations not at stoplights have a more scatter distribution and a higher range of prices. This could support the claim. Based on a bootstrap analysis of gas prices, a 95% confidence interval was constructed to estimate the true difference in mean prices between gas prices with and without direct competitors. The results indicate that we can be 95% confident that the true mean difference in gas prices falls between **-0.0554 and 0.0079**. The bootstrapped mean estimate of **-0.0235** suggests that, on average, gas stations with no competitors in sight tend to have slightly lower prices than those with competitors in sight- but this difference is very small.

Conclusion:

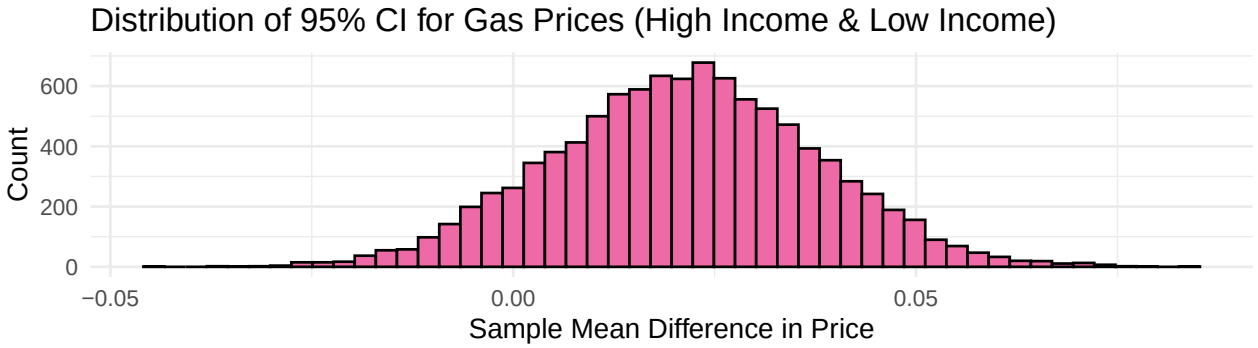
Since the confidence interval includes 0, the data does not provide strong evidence that competition significantly impacts gas prices. The bootstrap distribution appears to be approximately normal, supporting the reliability of our bootstrap. However, given the interval's range, we cannot conclude with statistical certainty that a meaningful difference in price exists between these two groups.

B)

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

Evidence:


```
##       name       lower     upper level     method   estimate
## 1 diffmean -0.01053999 0.0529868  0.95 percentile 0.02132411
```



Before conducting the bootstrap, I made a scatter plot to visualize the relationship between gas prices and income. The calculated correlation coefficient, 0.40, is positive and moderately strong, which would appear to support the claim. Based on a bootstrap analysis of gas prices, a 95% confidence interval was constructed to estimate the true difference in mean prices between low income and high income areas. These two variables were defined by calculating the median income in the data set, which was **$52,306**. The results indicate that we can be 95% confident that the true mean difference in gas prices falls between **-0.011 and 0.0534**. The bootstrapped mean estimate of **0.0213** suggests that, on average, gas stations that are in high income areas tend to have slightly higher prices than those located in low income areas- but the difference is very small.

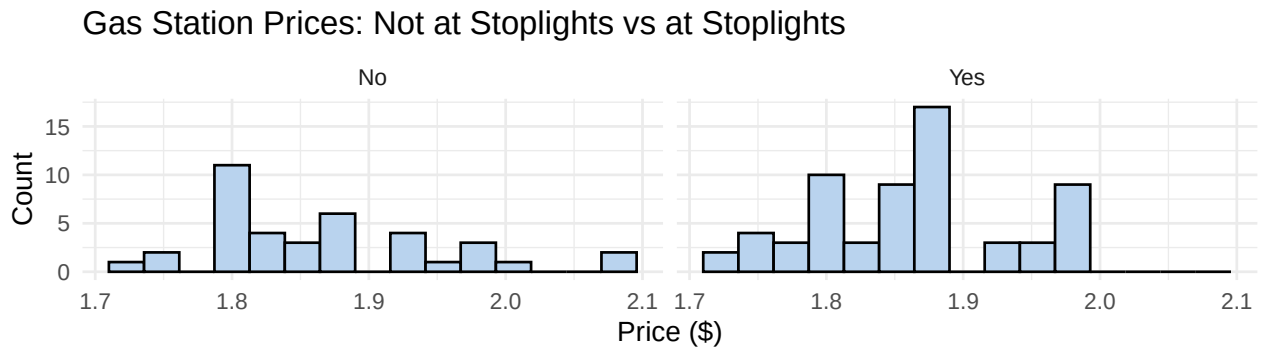
Conclusion:

Since the confidence interval includes 0, the data does not provide strong evidence that being in a high income area significantly impacts gas prices. The bootstrap distribution appears to be approximately normal, supporting the reliability of our bootstrap. However, given the interval's range, we cannot conclude with statistical certainty that a meaningful difference in price exists between these two groups.

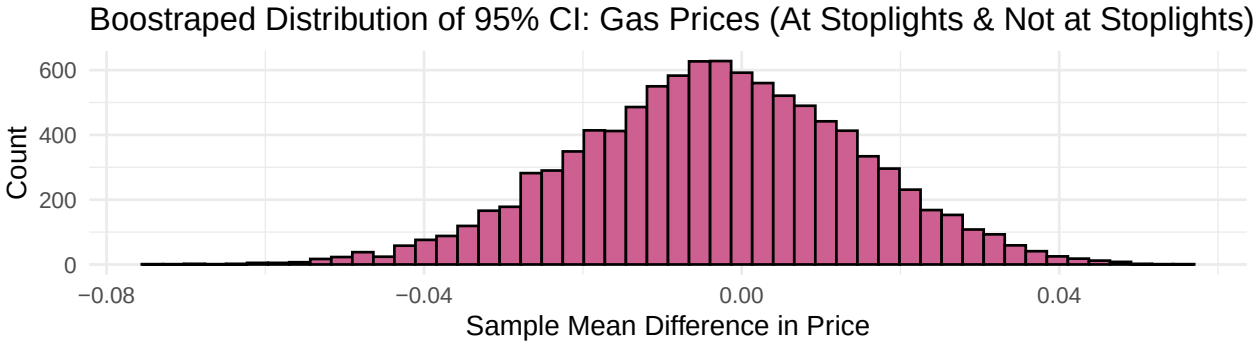
C)

Claim: Gas stations at stoplights charge more.

Evidence:




```
##       name       lower     upper level     method     estimate
## 1 diffmean -0.03887391 0.0306663  0.95 percentile -0.003299916
```



Before conducting the bootstrap, I made a graph to visualize the difference between gas prices between gas stations at stoplights and gas stations not at stoplights. Just based on these graphs, it appears that stations not at stoplights have a more scattered distribution, which stretches to a higher price than stations at stoplights. This appears to go against the claim. Based on a bootstrap analysis of gas prices, a 95% confidence interval was constructed to estimate the true difference in mean prices between gas stations at stoplights and gas stations not at stoplights. The results indicate that we can be 95% confident that the true mean difference in gas prices falls between **-0.0382 dollars and 0.0308 dollars**. The bootstrapped mean estimate is **-0.0031 dollars**.

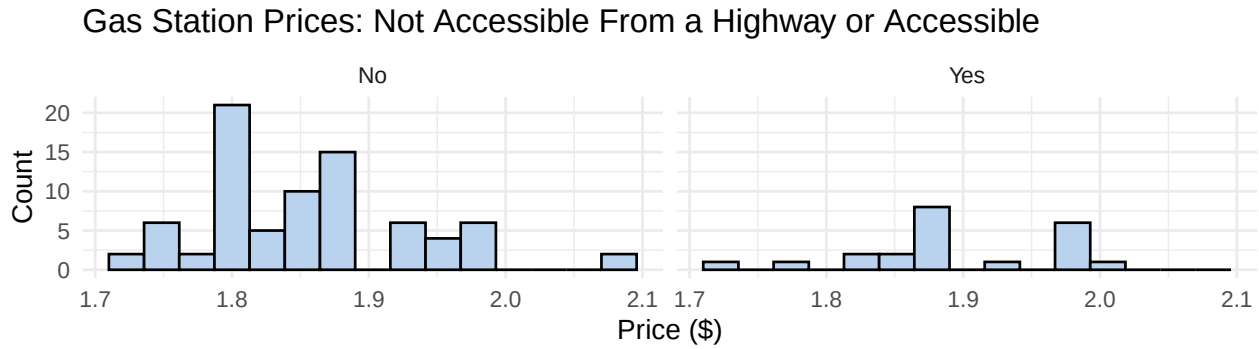
Conclusion:

Since the confidence interval includes 0, the data does not provide strong evidence that being at a stoplight significantly impacts gas prices. The bootstrap distribution appears to be approximately normal, supporting the reliability of our bootstrap. However, given the interval's range, we cannot conclude with statistical certainty that a meaningful difference in price exists between these two groups.

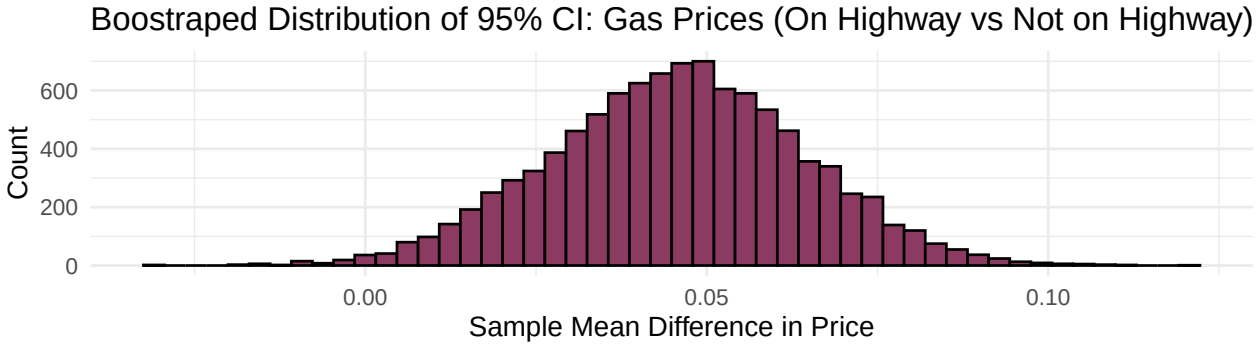
D)

Claim: Gas stations with direct highway access charge more.

Evidence:




```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008722511 0.08151873  0.95 percentile 0.0456962
```



The graphs show gas station prices for stations that are not accessible from either a highway or a highway access road versus those that are accessible. The distributions for non-accessible stations is moderately right-skewed, with a wide range of prices. Stations that are accessible however, is not skewed and smaller range of prices. Based on a bootstrap analysis of gas prices, a 95% confidence interval was constructed to estimate the true difference in mean prices between gas stations without direct high way access and gas stations with no direct highway access. The results indicate that we can be 95% confident that the true mean difference in gas prices falls between **0.0083 dollars and 0.0807 dollars**. The bootstrapped mean estimate is **0.0457 dollars**.

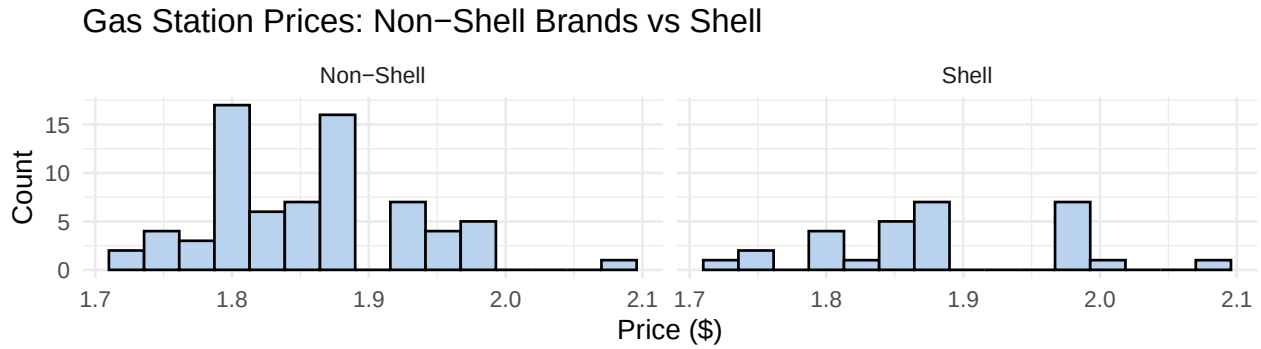
Conclusion:

Since the confidence interval does not include 0, the data provides evidence that gas stations with direct highway access have slightly higher prices than those without highway access. The bootstrap distribution appears to be approximately normal, supporting the reliability of our bootstrap. Given the interval's range, we can conclude with 95% certainty that a meaningful difference exists between these two groups.

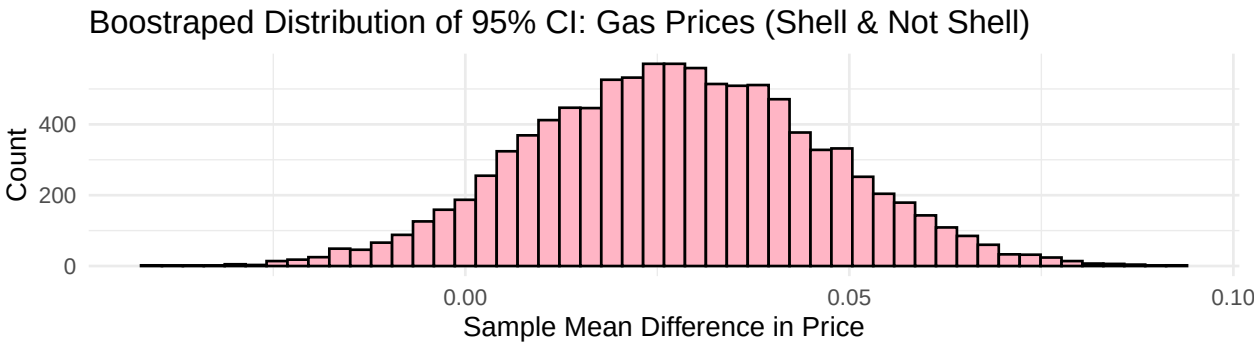
E)

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

Evidence:




```
##       name        lower      upper level     method   estimate
## 1 diffmean -0.008881385 0.06454874  0.95 percentile 0.02740421
```



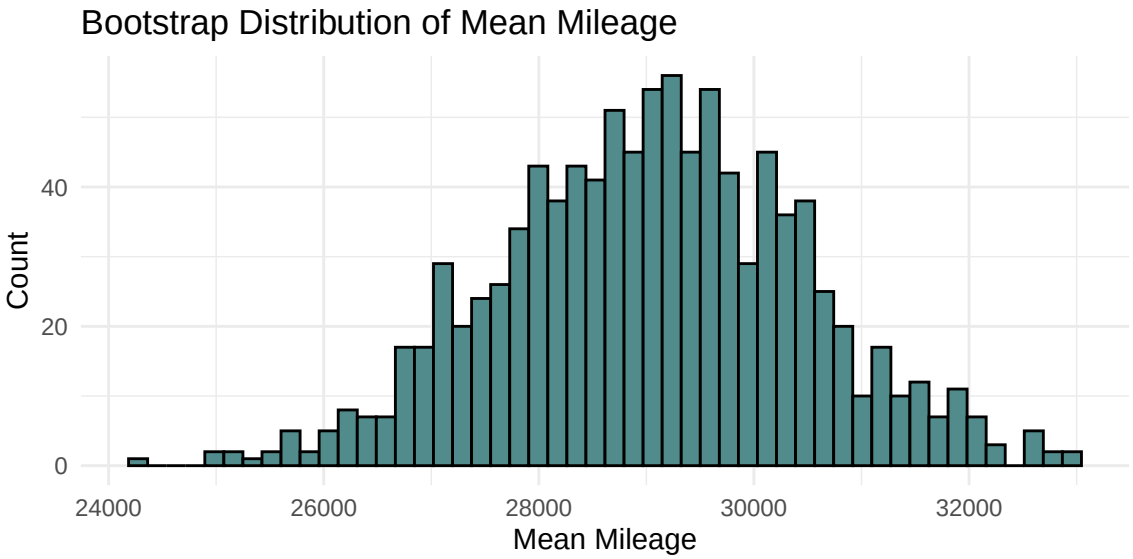
These graphs show gas station prices for non-Shell brands versus Shell. The distribution for non-Shell brands is slightly skewed-right compared to the fairly scattered distribution of Shell stations. Based on a bootstrap analysis of gas prices, a 95% confidence interval was constructed to estimate the true difference in mean prices between Non-Shell gas stations and gas stations with Shell gas stations. The results indicate that we can be 95% confident that the true mean difference in gas prices falls between **-0.0097 dollars and 0.0652 dollars**. The bootstrapped mean estimate is **0.0274 dollars**.

Conclusion:

Since the confidence interval includes 0, the data does not provide strong evidence that Shell gas stations charge more than non-Shell gas stations. The bootstrap distribution appears to just slightly skewed right. So, given the interval's range, we cannot conclude with statistical certainty that a meaningful difference in price exists between these two groups.

Problem 2

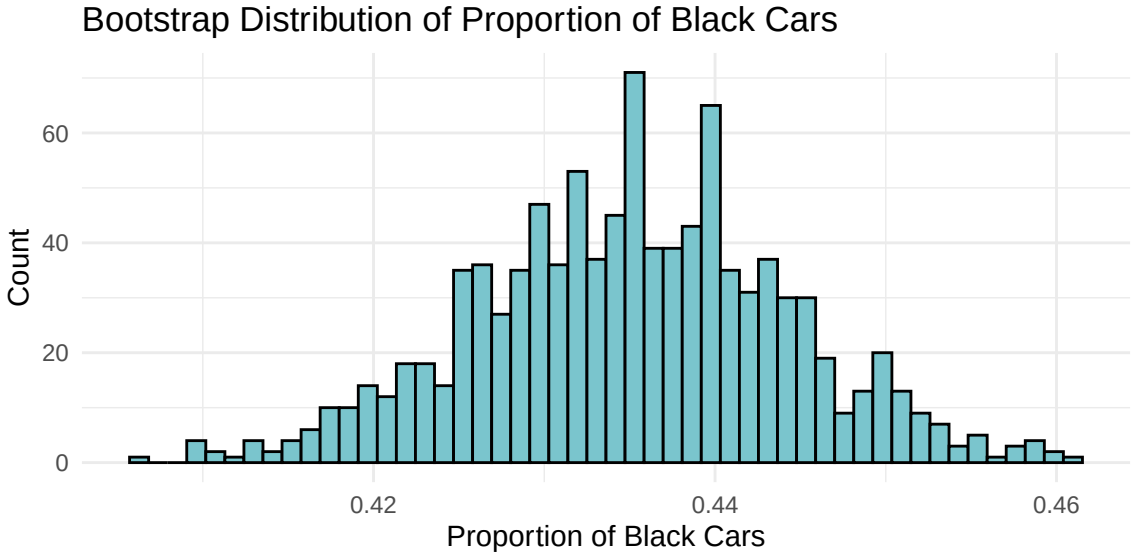
Part A:




```
##   name    lower    upper level     method estimate
## 1 mean 26247.01 31871.14  0.95 percentile 28997.34
```

Based on the analysis of 116 used 2011 S-Class 63 AMG vehicle, a 95% bootstrap confidence interval was constructed to estimate the true average mileage of these cars in the used-car market. The results indicate that we can be 95% confident that the true mean mileage falls between **26,2931.6 miles and 31,846.31 miles**, with a bootstrapped mean estimate of **28,997.34 miles**. The bootstrap standard error of **1,430.64 miles** quantifies the variability in our estimate, indicating a moderate spread in the sampled means. The histogram of the bootstrap distribution suggests an approximately normal shape, supporting the reliability of the bootstrap approach. Overall, this confidence interval provides a strong statistical basis for estimating the expected mileage range of 2011 S-Class 63 AMG vehicles available in the used-car market at the time of data collection.

Part B:




```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4170993 0.4527605  0.95 percentile 0.4347525
```

Based on a sample of 2,889 used 2014 S-Class 550 vehicles, a 95% bootstrap confidence interval was constructed to estimate the true proportion of these cars that were painted black. The analysis yielded a bootstrap mean estimate of **0.4348 (or 43.48%)**, with a 95% interval ranging from **41.64% to 45.31%**. This suggests that we can be 95% confident that the true proportion of black 2014 S-Class 550s in the used-car market falls within this interval. The bootstrap histogram shows that the distribution of bootstrapped proportions is approximately normal, which supports the reliability of this estimation method. Additionally, calculating the bootstrap standard error, **0.0091** helps quantifies the variability in these estimates. Overall, this confidence interval provides a strong statistical basis for estimating the prevalence of black-painted 2014 S-Class 550 vehicles in the used-car market at the time of data collection.

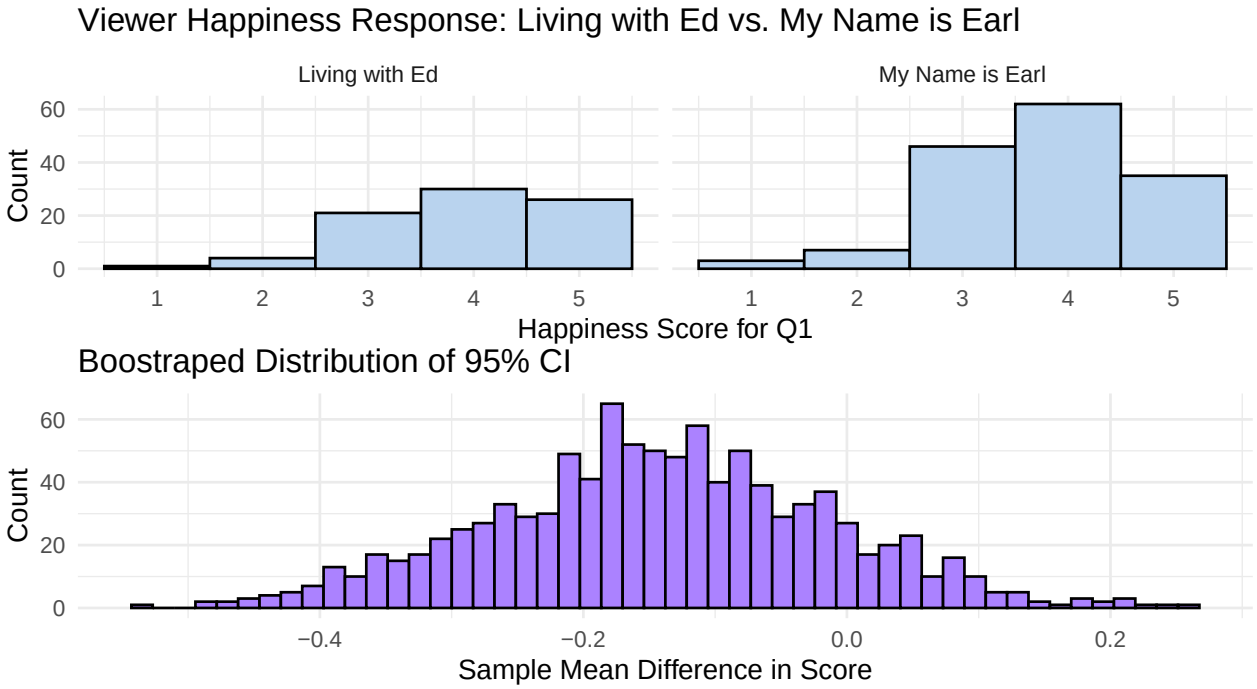
Problem 3

Part A:

Question: Who makes people happier: Ed or Earl?

Approach: To try to answer this question and determine if there is evidence that one show consistently produces a higher mean Q1_Happy response among viewers, I made 2 graphs comparing the distributions and conducted a 95% confidence interval bootstrap.

Results:




```
##       name      lower      upper level     method   estimate
## 1 diffmean -0.3943521 0.09920766  0.95 percentile -0.1490515
```

Conclusion:

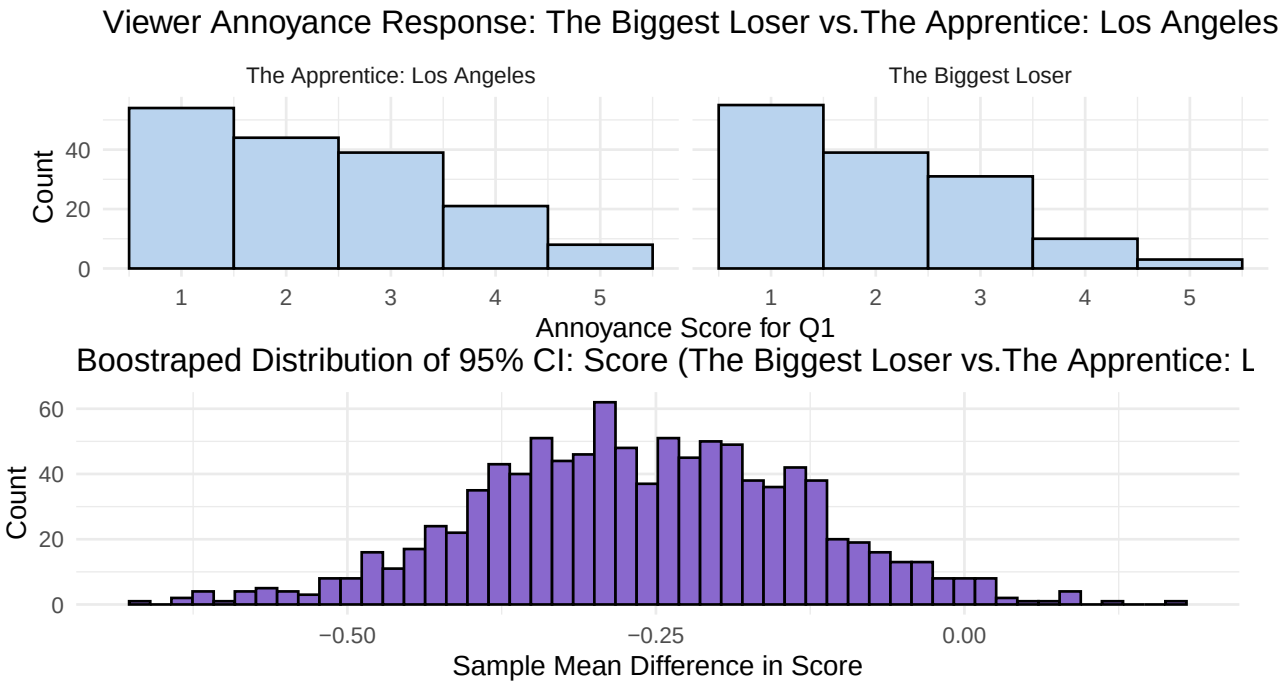
The histogram shows similar distributions of happiness scores for both shows, with Living with Ed having a slight concentration of higher ratings. The mean estimate of this confidence interval is **-0.1491.** Based on a bootstrap analysis of gas prices we can be 95% confident that the true mean difference in Q1_Happy scores between Living with Ed and My Name is Earl falls between **-0.401 to 0.106**. Since the confidence interval includes zero, the statistical evidence does not confirm a consistent difference- and we cannot conclusively say that either show makes viewers significantly happier than the other.

Part B:

Question: Consider the shows "The Biggest Loser" and "The Apprentice: Los Angeles." Which reality/contest show made people feel more annoyed?

Approach: To try to answer this question and determine if there is evidence that one show consistently produces a higher mean Q1_Annoyed response among viewers, I conducted a bootstrap with a 95% confidence interval

Results:




```
##       name      lower        upper level     method  estimate
## 1 diffmean -0.5180814 -0.006980701  0.95 percentile -0.270997
```

Conclusion:

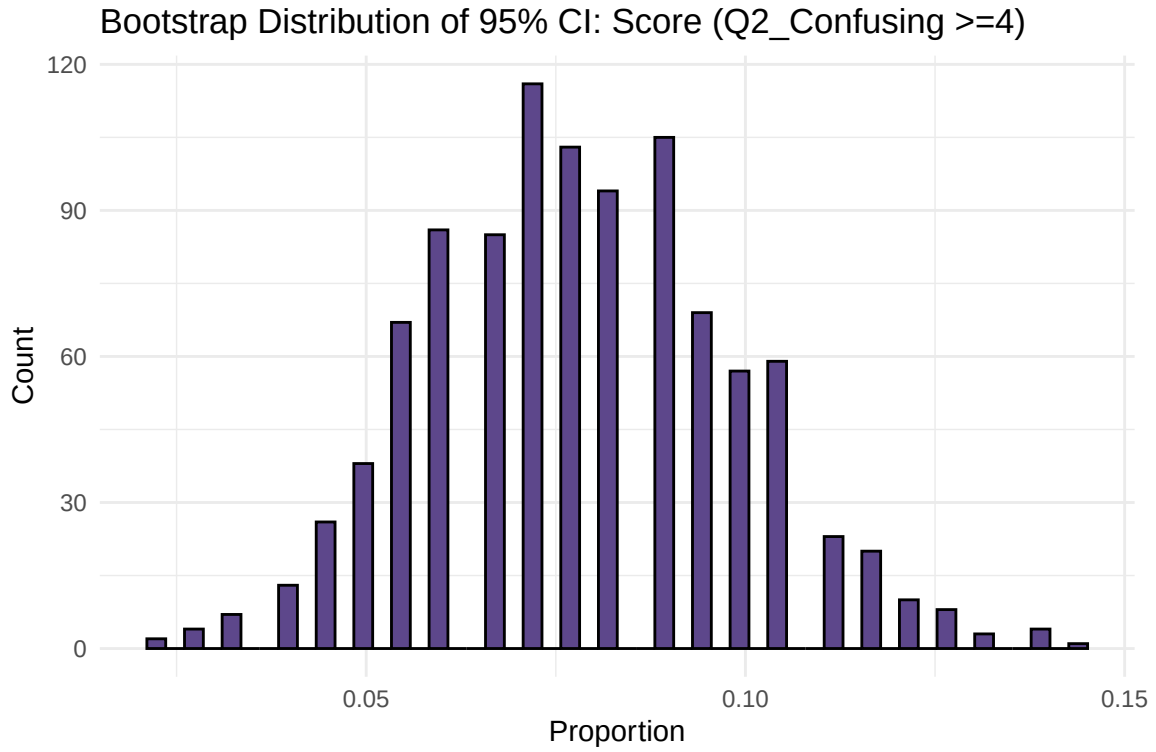
The histogram shows a higher concentration of lower annoyance scores for *The Biggest Loser*, while The Apprentice: Los Angeles has a wider distribution, including more higher scores. The mean estimate of this confidence interval is -**0.271.** Based on a bootstrap analysis of gas prices we can be 95% confident that the true mean difference in Q1_Annoyed scores betweenhe Apprentice: Los Angeles and he Biggest Loser falls between -**0.513 to -0.0304**. Since the confidence interval does not include zero, there is an indication of a statistically significant difference at the 95% confidence level. While this is convincing and technically significant level, it would be helpful to do further research.

Part C:

Question: Based on this sample of respondents, what proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question?

Approach: To try to answer this question and determine the hat proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question, I did a bootstrap with a 95% confidence interval.

Results:




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

Conclusion:

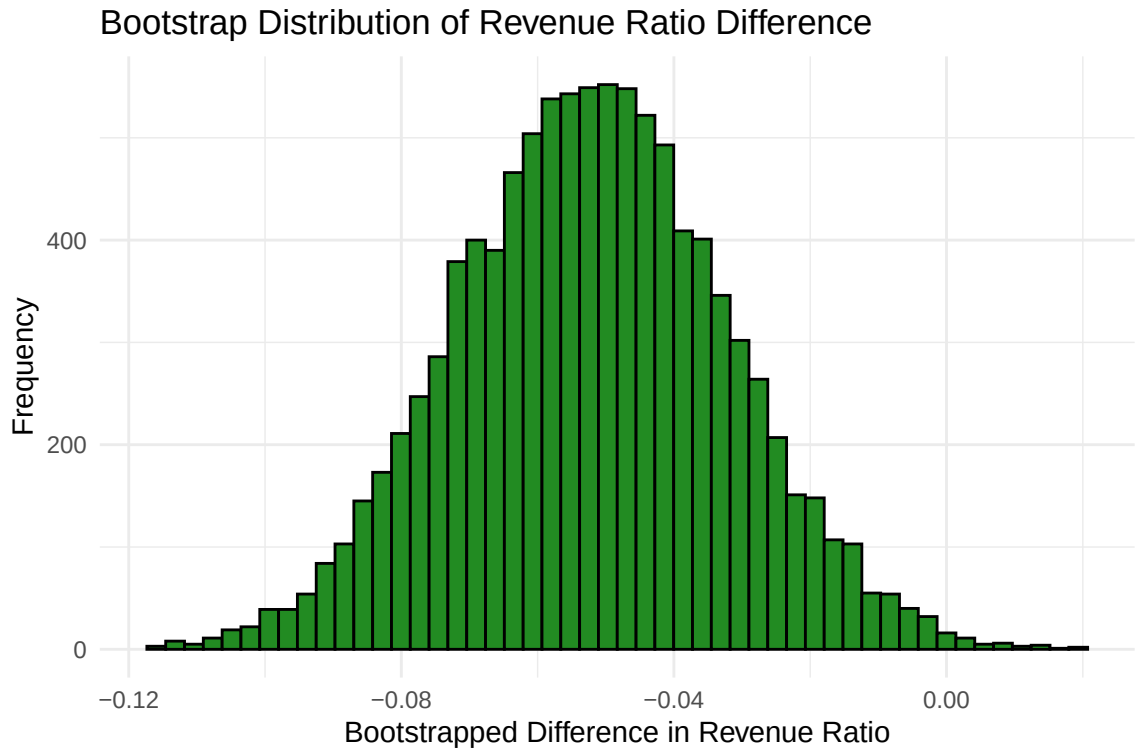
Based on this sample of respondents, we are 95% confident that the true proportion of American TV watchers who would give a response of **4 or greater** to the Q2_Confusing question falls between **4.42% and 12.15%**, with an estimate of **7.73%**. This suggests that while the majority of viewers do not find the show confusing, a small but notable percentage would rate it as such. The bootstrap distribution appears approximately normal, indicating that our resampling approach provides a reliable estimate of this proportion.

Problem 4:

Question: What is the difference in revenue ratio between the treatment (adwords paused) and control (adwords running) DMAs? Does the evidence suggest that the revenue ratio is the same for both groups, or does it support the idea that paid search advertising on Google generates additional revenue for EBay?

Approach: I computed the revenue ratio for each DMA (rev_after / rev_before) and calculated the difference in mean revenue ratio between the treatment and control groups. To assess statistical significance, I used a bootstrap analysis with 10,000 Monte Carlo simulations to generate a confidence interval for this difference.

Results:




```
## [1] "95% Confidence Interval: -0.0905 to -0.0131"
```

```
## [1] "Observed Difference in Revenue Ratio: -0.0523"
```

The observed difference in revenue ratio between the treatment and control groups is **-0.0523**, indicating that DMAs where paid search advertising was paused experienced a lower revenue ratio compared to those where ads remained active. To assess the statistical significance of this difference, I conducted a bootstrap analysis with 10,000 Monte Carlo simulations, generating a 95% confidence interval of **-0.0899 to 0.0132.** Since this confidence interval does **not include zero**, it suggests that the difference in revenue ratio is statistically significant. This means that the revenue ratio in the treatment group is **consistently lower** than in the control group, providing evidence that pausing paid search ads negatively impacted revenue.

Conclusion:

The data suggests that pausing paid search ads leads to a **statistically significant decrease** in revenue. Since the confidence interval is entirely negative, this indicates that Google ads are likely

contributing to increased revenue for EBay. This supports the conclusion that paid search advertising is **not redundant** and plays a role in driving sales.