HW3

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
##
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
##
## Attaching package: 'mosaic'
##
##
## The following object is masked from 'package:Matrix':
##
##     mean
##
##
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
##
##
## The following object is masked from 'package:purrr':
##
##     cross
##
##
## The following object is masked from 'package:ggplot2':
##
##     stat
##
##
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
##
##
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

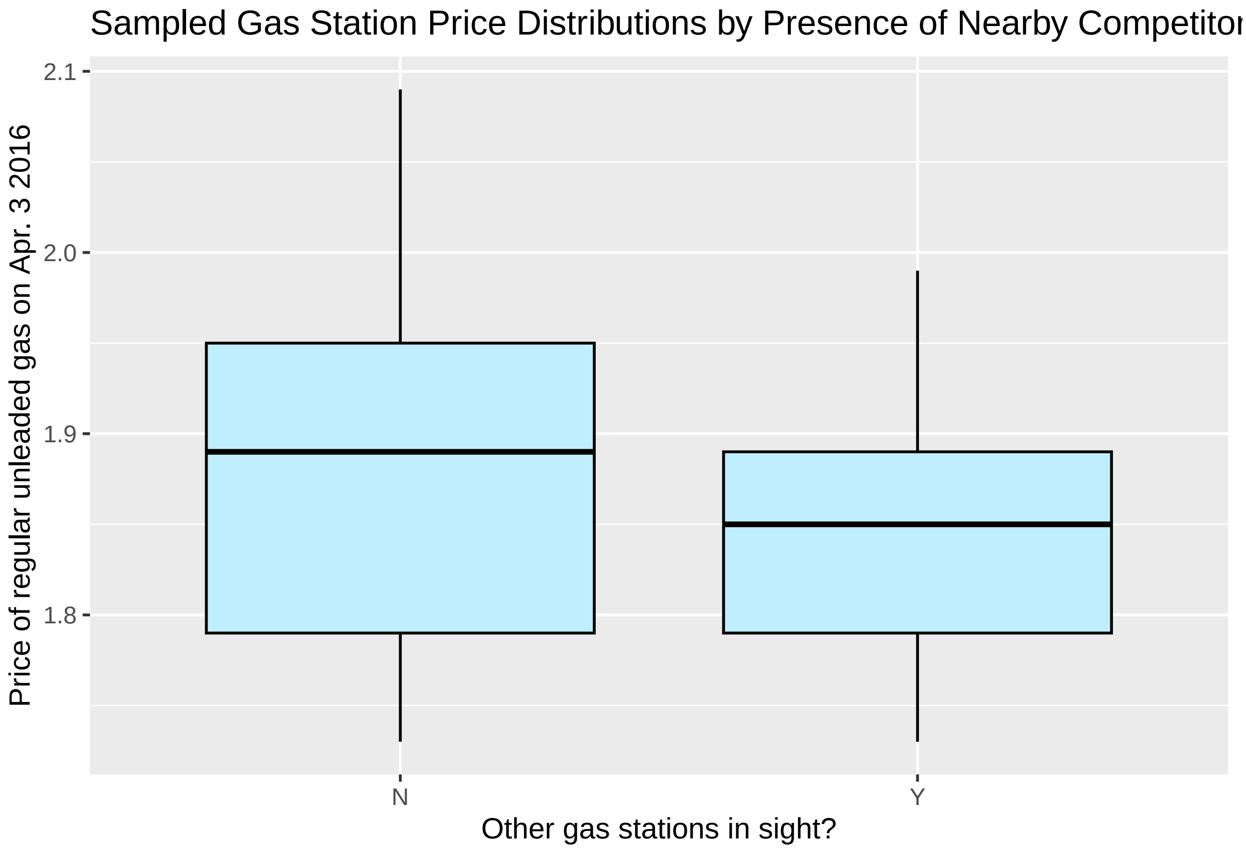
Problem 1

Theory A

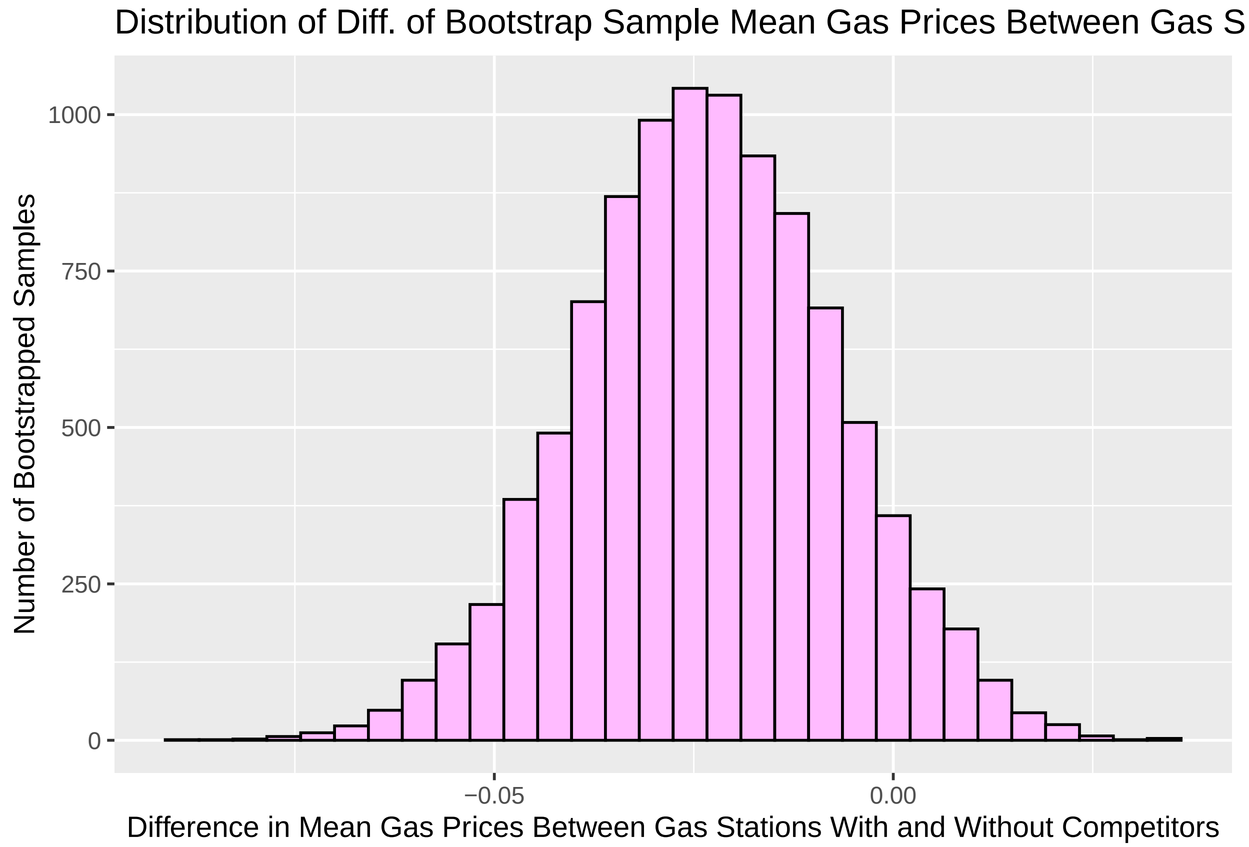
Claim

Theory A states that “gas stations will charge more if they lack direct competition in sight”.

Evidence



Across the 101 gas stations in the sample, gas stations with no competitors in sight had their median gas price on April 3 2016 about 0.04 dollars higher than gas stations with competitors in sight.



```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05538575 0.008538038  0.95 percentile -0.03630687
```

10000 bootstrapped samples were constructed from the original sample of 101 gas stations, and in each bootstrap sample the difference in mean gas station prices between gas stations without and with competitors in sight was taken. The distribution of the differences in means is displayed in the histogram above. A 95% confidence interval of the difference in means was calculated; based on this, with 95% confidence, the difference in mean prices between gas stations without and with competitors in sight is between -0.0553 and 0.00853 dollars.

Conclusion

The 95% confidence interval of difference in mean gas prices between gas stations without and with competitors captures 0, so the difference in means is not statistically significant at the 5% level. Therefore, there is not enough evidence to show that gas stations without nearby gas stations in sight charge higher gas prices than gas stations with nearby gas stations in sight.

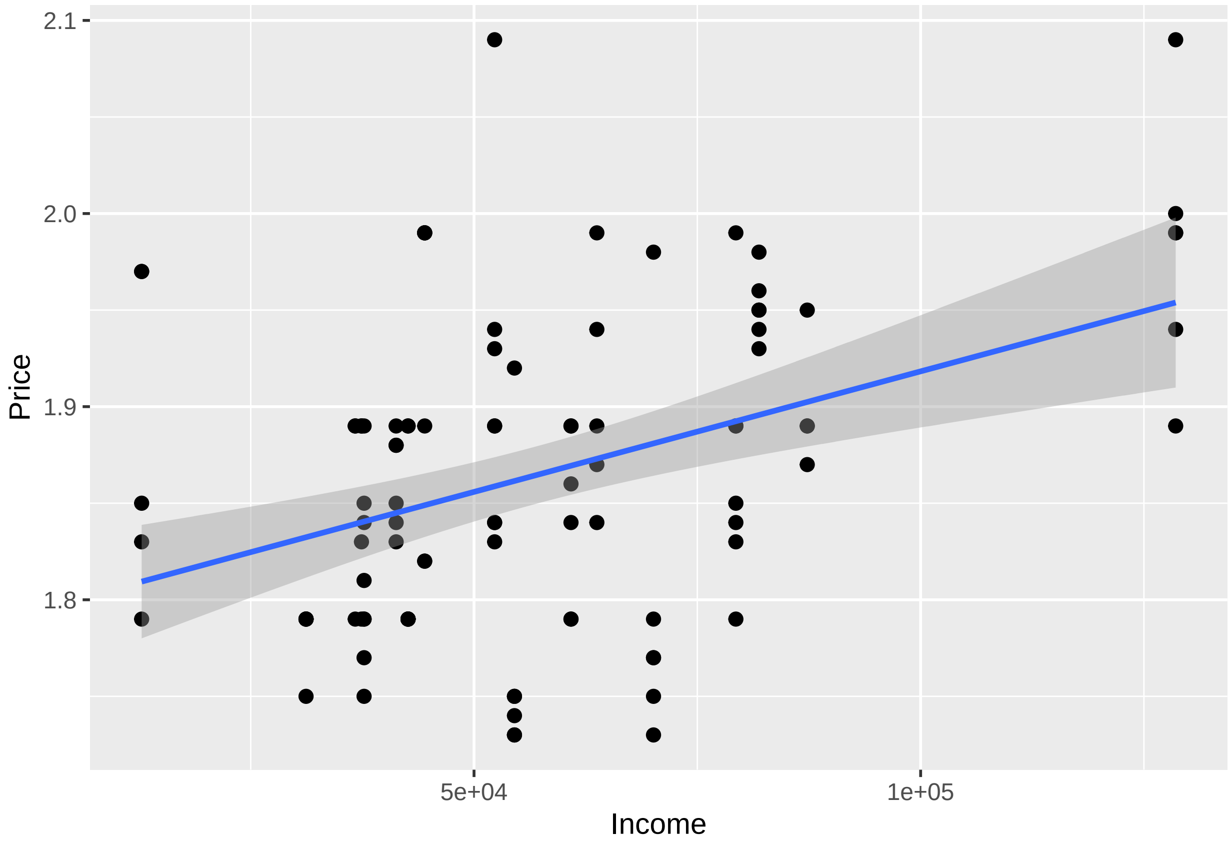
Theory B

Claim

Theory B claims that richer areas have higher gas prices.

Evidence

First, it would be helpful to present a scatterplot demonstrating the relationship between gas prices and income within the sample of 101 gas stations.

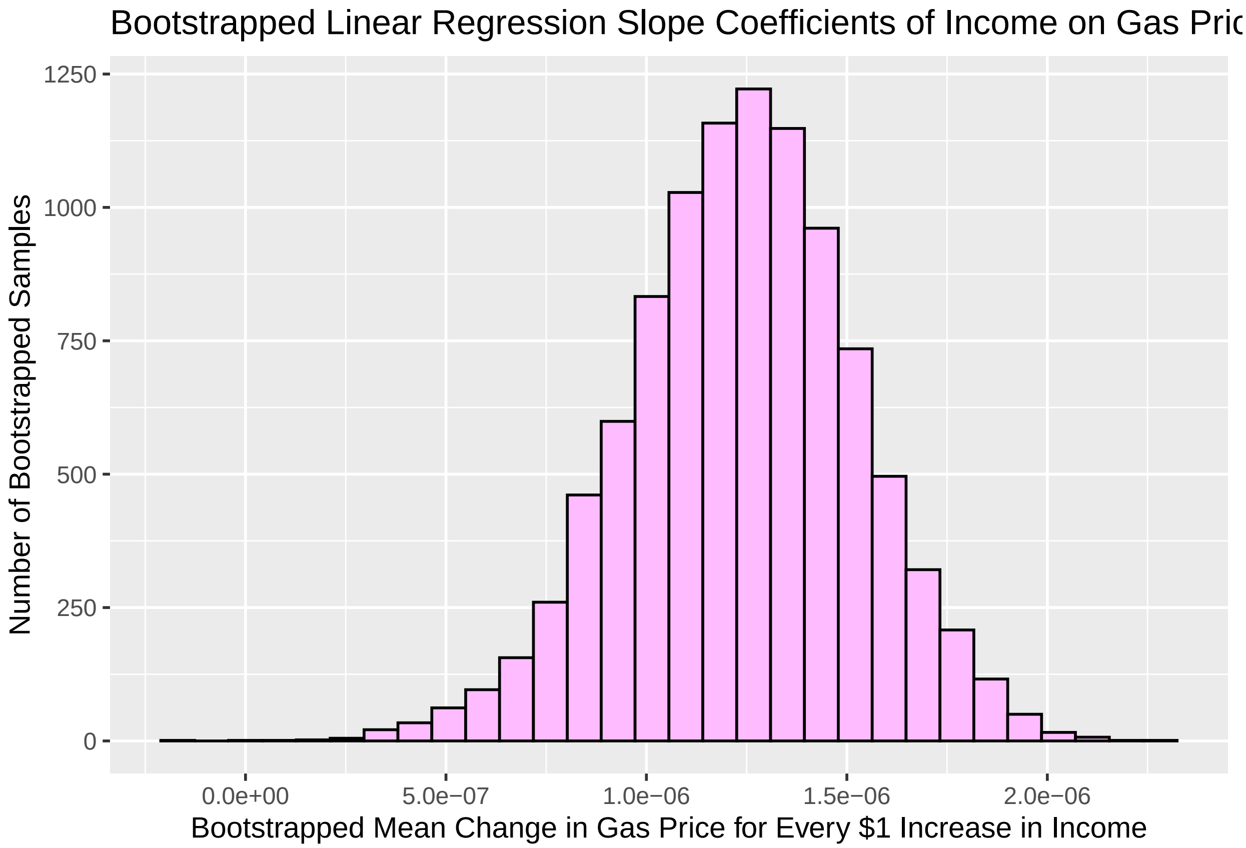


```
## $x
## [1] "2014 Median Household Income by Zip Code of Gas Stations"
##
## $y
## [1] "Price of Regular Unleaded Gas on Apr. 3 2016"
##
## $title
## [1] "Relationship Between Income and Gas Prices across 101 Sampled Gas Stations"
##
## attr(,"class")
## [1] "labels"
```

```
##
## Call:
## lm(formula = Price ~ Income, data = gas_prices)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.150945 -0.050492 -0.009412  0.046076  0.231262
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.793e+00  1.811e-02  99.006  < 2e-16 ***
## Income      1.248e-06  2.908e-07   4.293 4.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.07522 on 99 degrees of freedom
## Multiple R-squared:  0.1569, Adjusted R-squared:  0.1484 
## F-statistic: 18.43 on 1 and 99 DF,  p-value: 4.117e-05
```

There appears to be a moderate, positive relationship between income and gas prices among the 101 sampled gas stations, as indicated by a Pearson correlation of 0.3962.



```
##        name        lower        upper level     method     estimate
## 1 Intercept 1.758729e+00 1.829458e+00  0.95 percentile 1.788903e+00
## 2    Income 6.513198e-07 1.787617e-06  0.95 percentile 1.340312e-06
## 3     sigma 6.366236e-02 8.501616e-02  0.95 percentile 6.230719e-02
## 4 r.squared 3.786234e-02 3.224546e-01  0.95 percentile 2.393912e-01
## 5         F 3.895879e+00 4.711568e+01  0.95 percentile 3.115889e+01
```

10000 bootstrapped samples were taken and a linear regression output was generated for each one, with income as the predictor and gas prices as the outcome variable. Above is the distribution of the bootstrapped linear regression slopes. A confidence interval was calculated from the bootstrapped samples, finding with 95% confidence that the mean change in gas prices for every one dollar increase in income lies between 0.0000006513 and 0.000001787764 dollars.

Conclusion

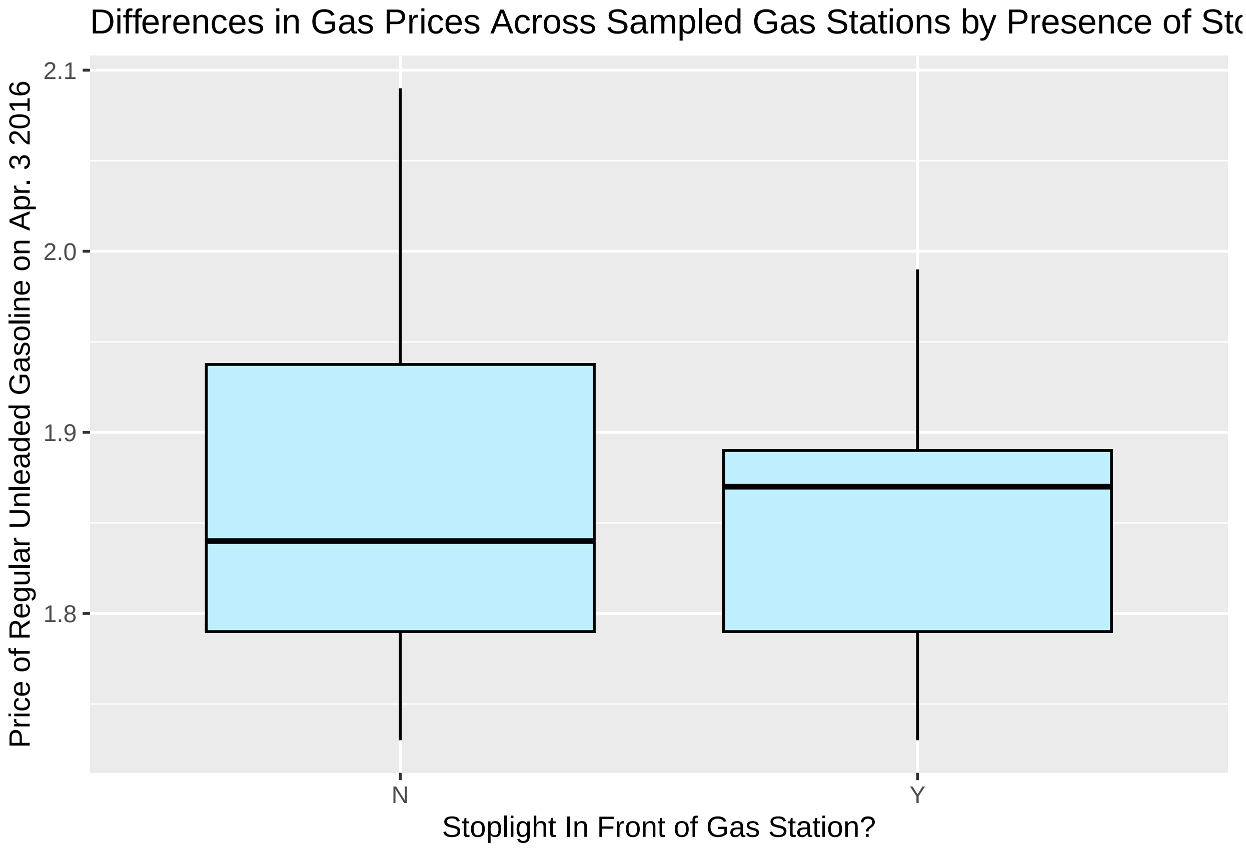
Since the 95% confidence interval of mean change of gas prices for every dollar increase in income does not capture 0, the data do appear to support the theory that richer areas have higher gas prices.

Theory C

Claim

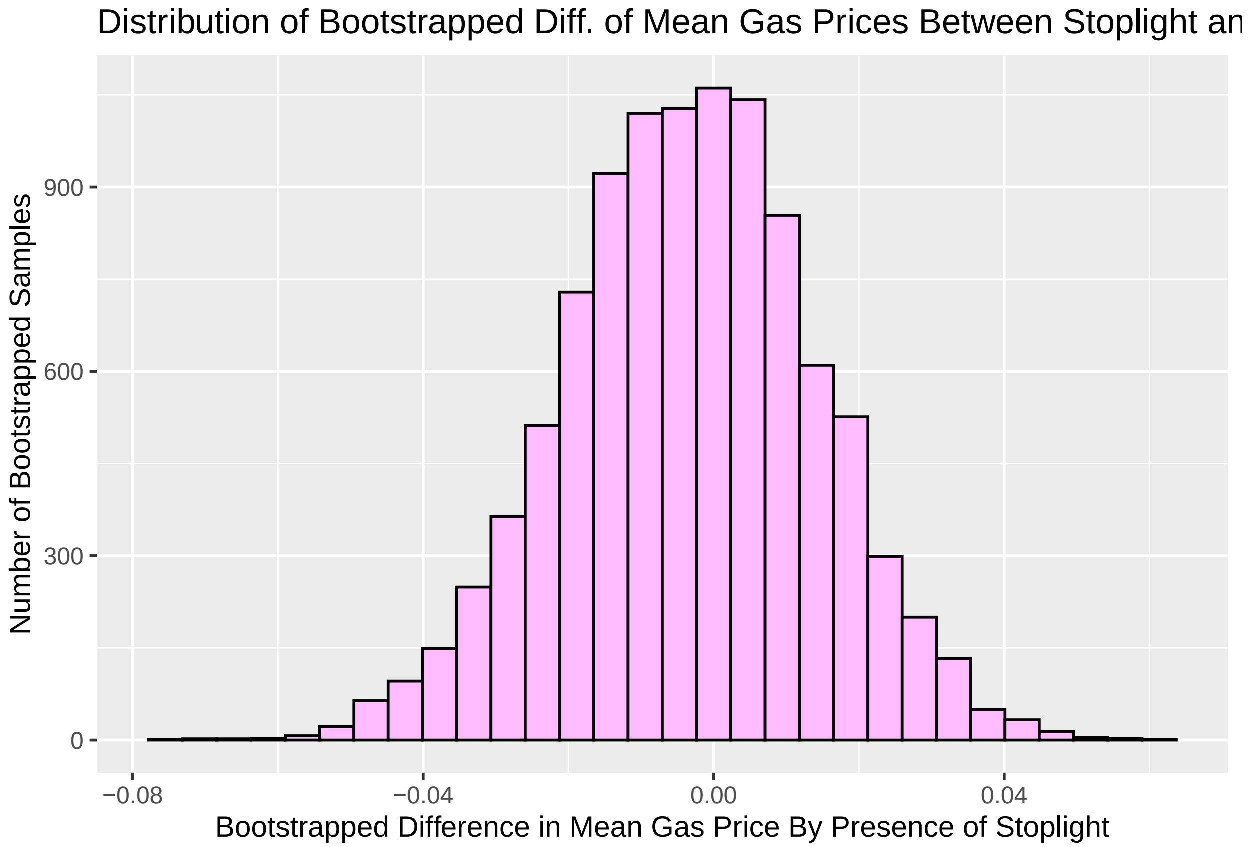
Theory C states that gas stations at stoplights have higher gas prices than gas stations not at stoplights.

Evidence



```
## [1] -0.0033
```

Sampled gas stations with a stoplight in front of them had a mean gas price around 0.0033 dollars lower than gas stations without stoplights in front of them.



```
##       name       lower     upper level     method      estimate
## 1 diffmean -0.03789405 0.0302501  0.95 percentile -0.0005653569
```

Bootstrap samples were generated from the original sample, and for each the difference in mean gas prices between gas stations with and without stoplights in front of them was calculated; this process was repeated 10,000 times and the output was stored in an object and the distribution of the bootstrapped diff. in means is shown in the histogram above. A confidence interval was generated from the 10000 bootstrapped diff. in means, determining with 95% confidence that the difference in mean gas prices between gas stations with and without a stoplight in front of them lies between -0.0378 and 0.0302 dollars.

Conclusion

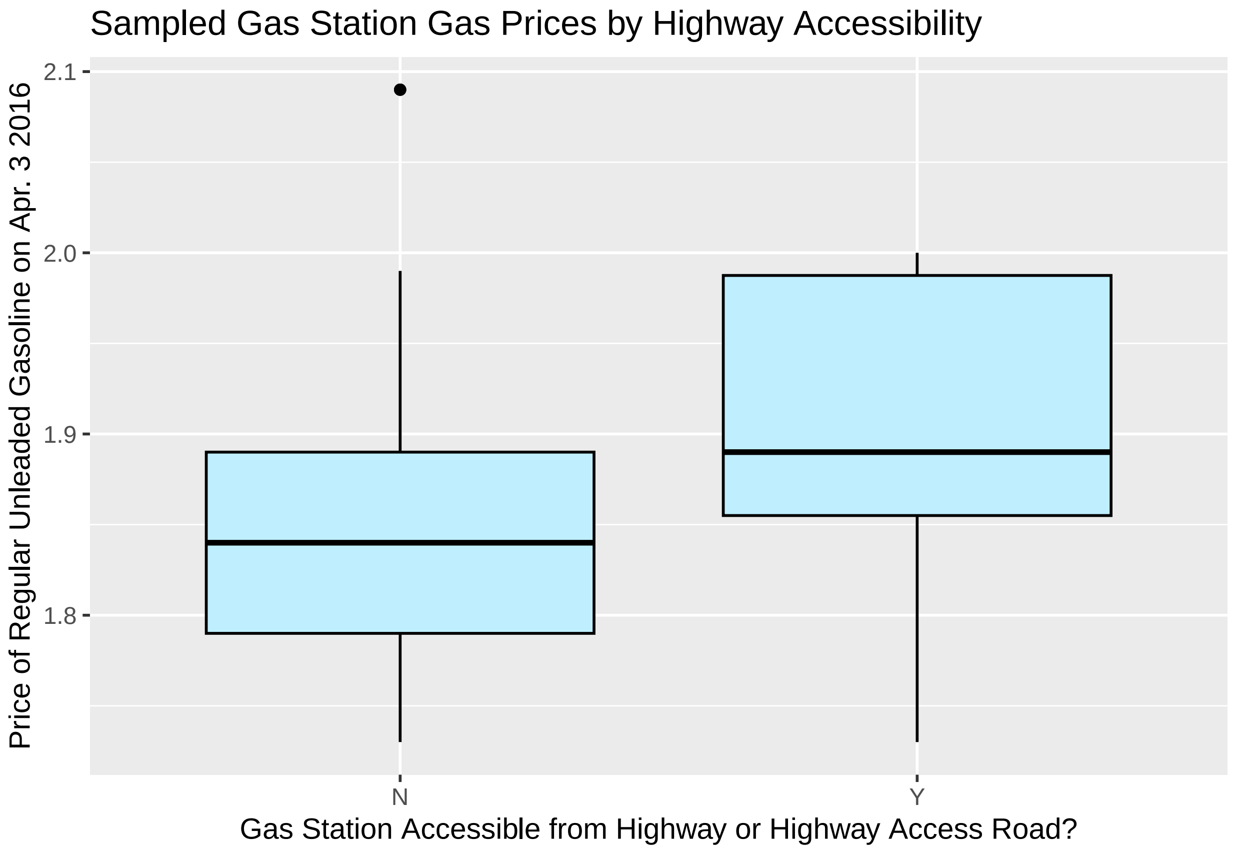
Since the 95% confidence interval of the difference in bootstrapped gas price means between gas statiosn with and without a stoplight in front of them captures 0, the data do not support the theory that gas stations with a stoplight in front of them have higher gas prices.

Theory D

Claim

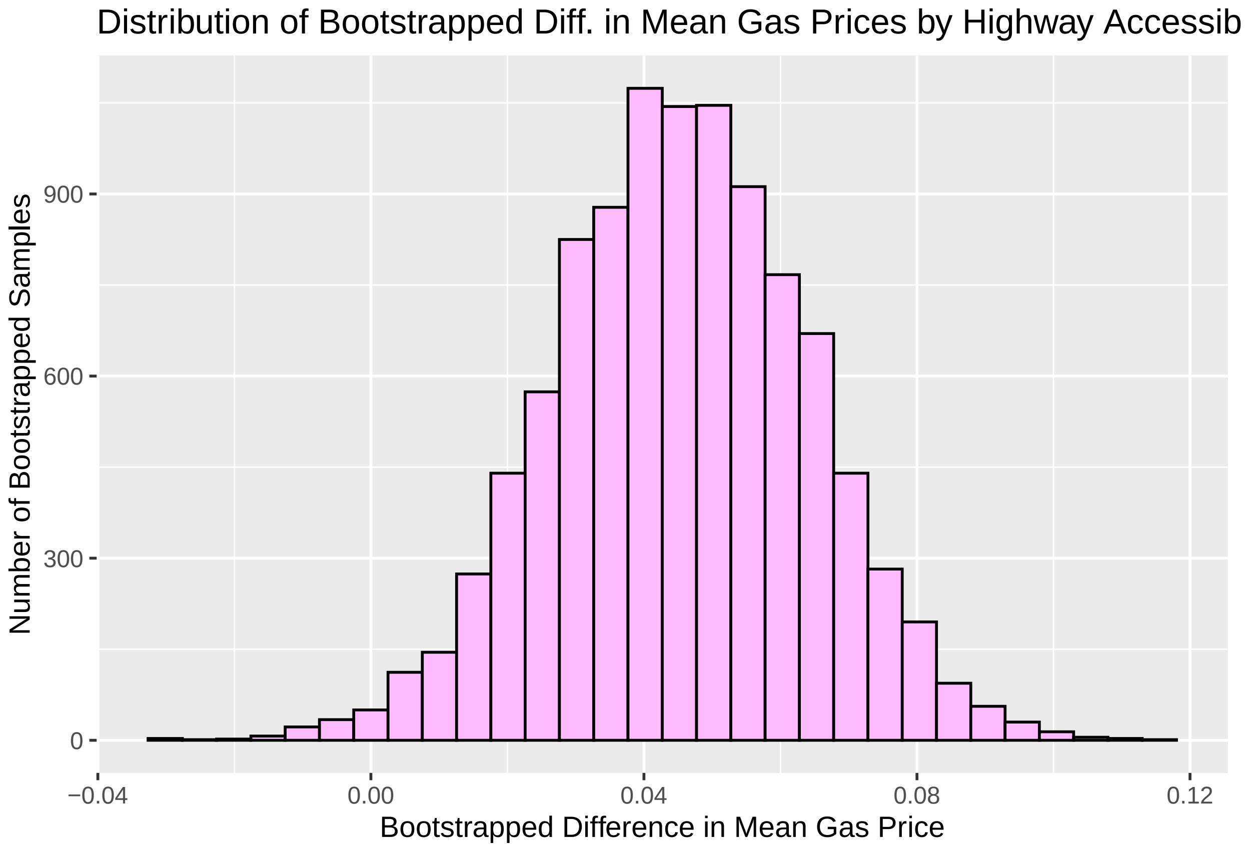
Theory D claims that gas stations accessible from a highway have higher gas prices than gas stations that are not accessible from a highway.

Evidence



```
## [1] 0.0457
```

Sampled gas stations that are accessible from a highway have gas prices 0.0457 dollars higher than those unaccessible from a highway.



```
##       name       lower     upper level     method   estimate
## 1 diffmean 0.008225616 0.0816305  0.95 percentile 0.04323276
```

Bootstrap samples were taken from the original sample, and the difference in mean gas prices between gas stations with and without highway accessibility was calculated; this process was repeated 10000 times and the output is displayed in the histogram distribution above. A confidence interval was calculated on the diff. in means, finding with 95% confidence that the difference in mean gas prices between gas stations with and without highway or highway access road accessibility lies between 0.00883 and 0.08134 dollars.

Conclusion

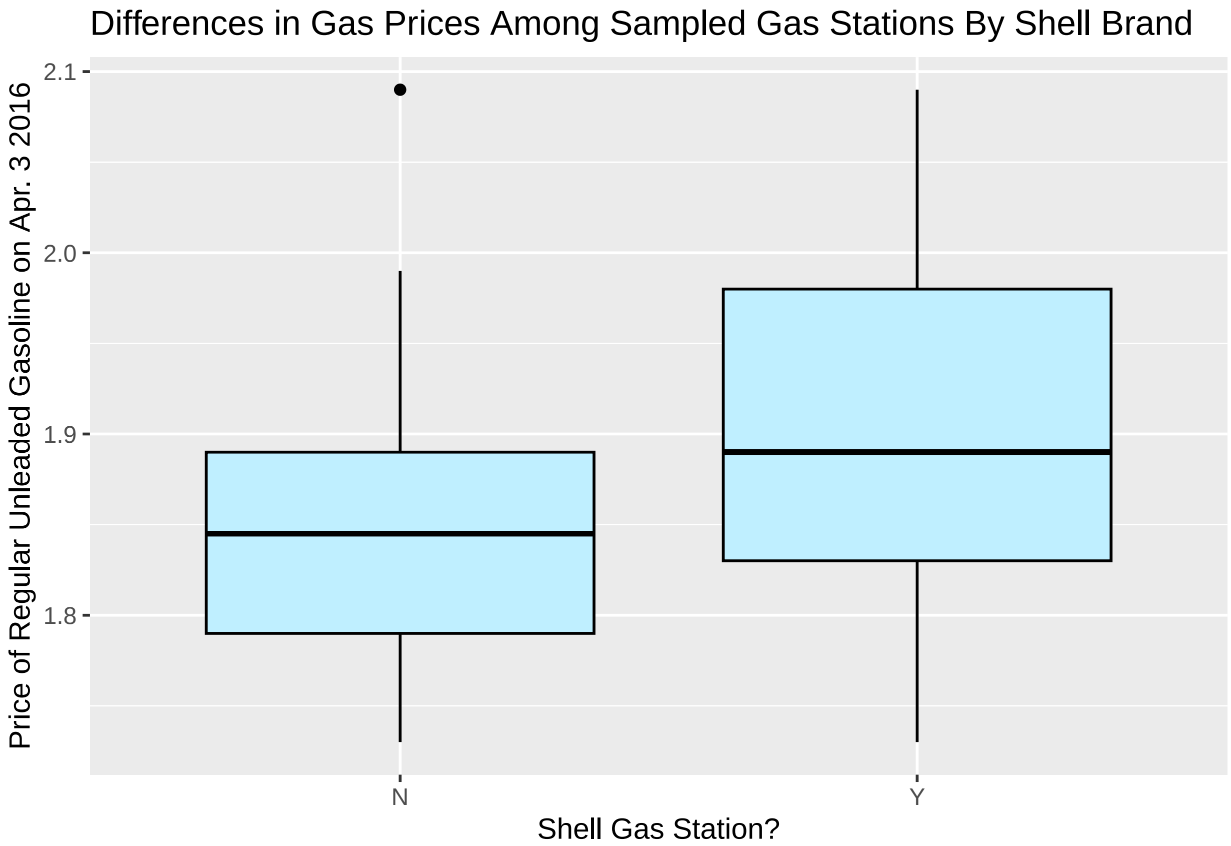
Since the 95% confidence interval of the difference in mean gas prices betwen gas stations with and without highway or highway access road accessbility does not capture 0, the data do support the theory that gas stations with direct highway access have higher gas prices.

Theory E

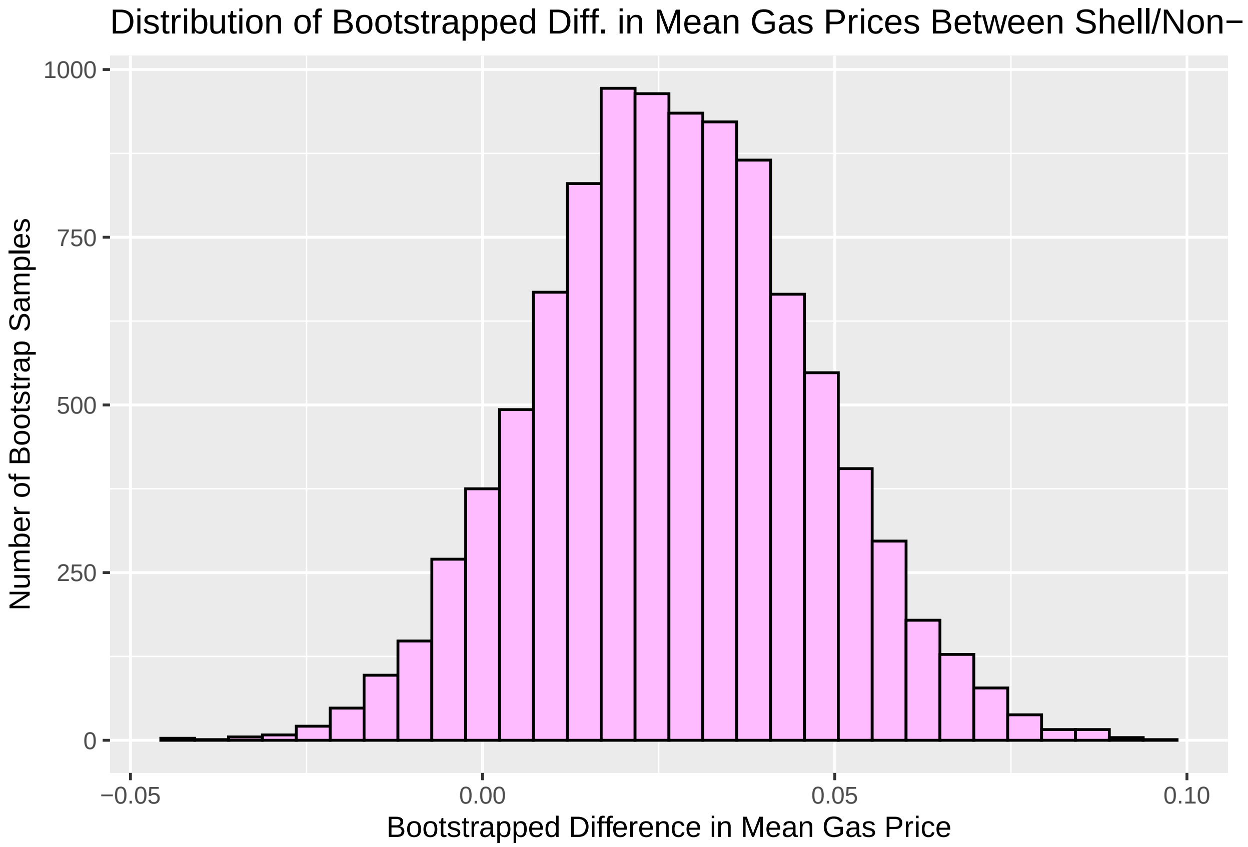
Claim

Theory E claims that Shell brand gas stations have higher gas prices than non Shell brand gas stations.

Evidence



Sampled Shell gas stations had a mean gas price 0.0274 dollars higher than non Shell gas stations.



```
##       name        lower      upper level     method   estimate
## 1 diffmean -0.009328656 0.06584863  0.95 percentile 0.03359877
```

Bootstrap samples were taken from the original sample and the difference in mean gas prices between Shell and non-Shell gas stations was calculated for each bootstrap sample, with the process repeated 10000 times and the distribution of difference in means displayed on the graph above. A confidence interval was produced, determinging with 95% confidence that the difference in mean gas prices between Shell and non-Shell gas stations lies between -0.00928 and 0.0653 dollars.

Conclusion

Since the 95% confidence interval of difference in mean gas prices between Shell and non-Shell gas stations captures 0, the data do not support the theory that Shell gas stations charge higher gas prices than non-Shell gas stations.

Problem 2

Part A

```
##   name    lower    upper level     method estimate
## 1 mean 26210.38 31821.11  0.95 percentile 24873.47
```

It can be stated with 95% confidence that the mean mileage of 2011 Mercedes S-Class 63 AMGs on cars.com at the time of data collection lies between 26223.03 and 31771.8 miles per gallon.

Part B

```
##        name     lower     upper level     method estimate
## 1 prop_TRUE 0.4167532 0.4534441  0.95 percentile 0.443406
```

It can be stated with 95% confidence that the proportion of 2014 Mercedes S-Class 550s on cars.com at the time of data collection that were black lies between 0.4168 and 0.4534.

Problem 3

Part A

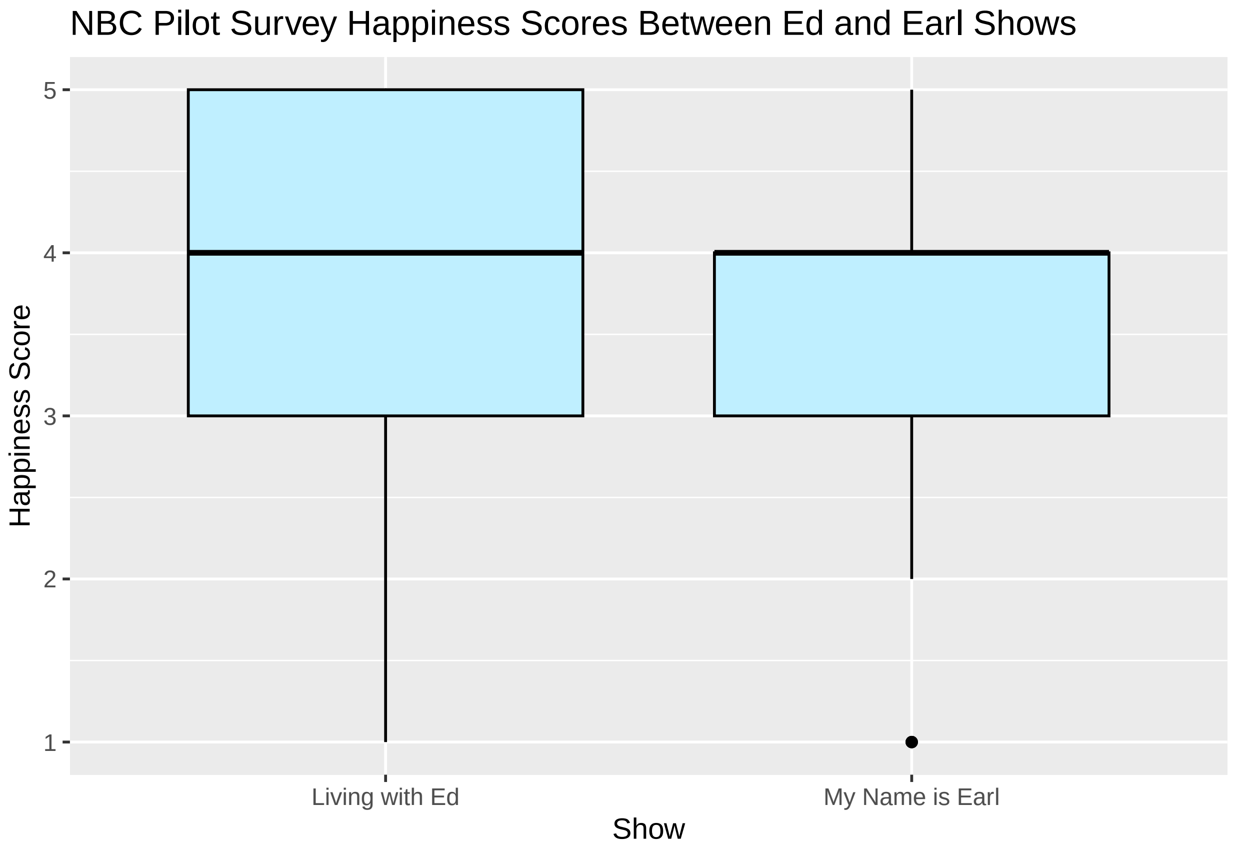
Question

What show makes people happier: "Living with Ed" or "My Name is Earl"?

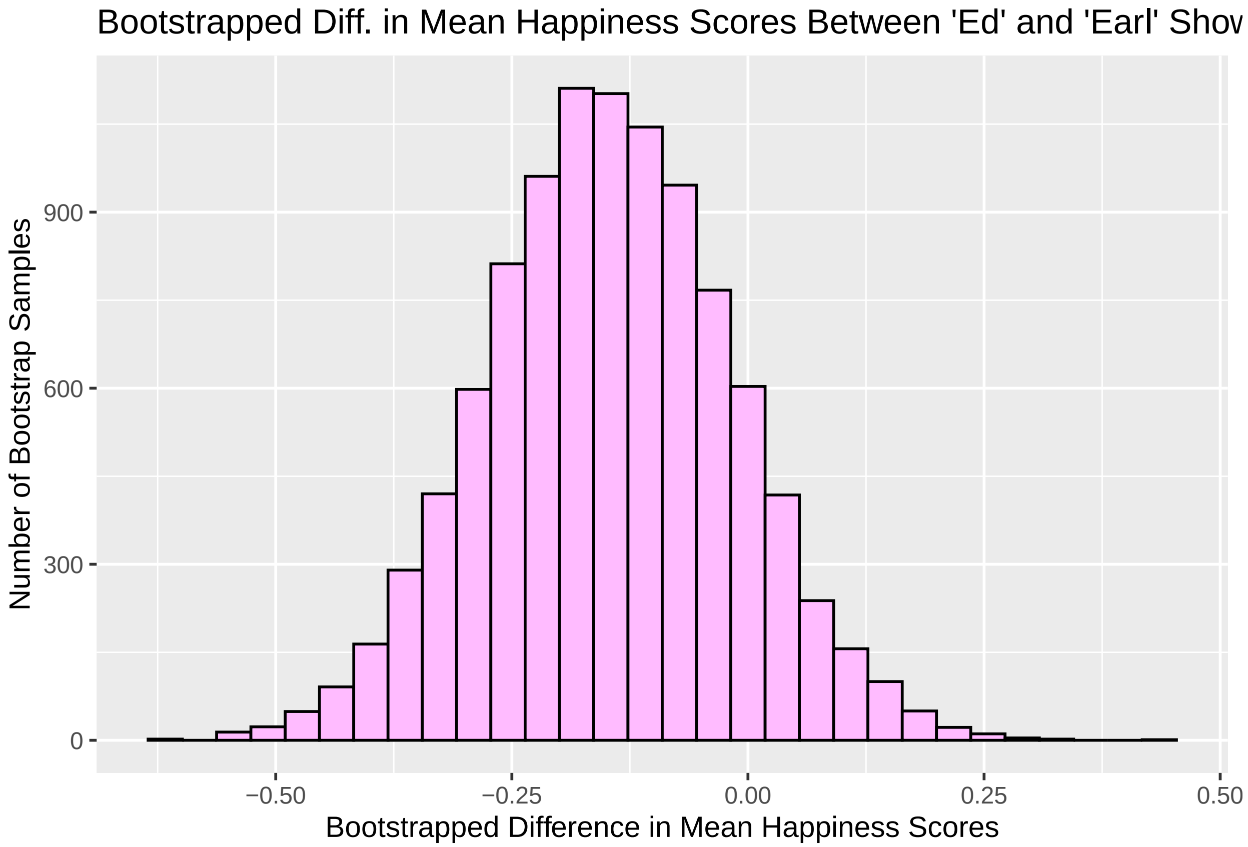
Approach

A dataset is created, only including observations on "Living with Ed" or "My Name is Earl", and to measure happiness, the `Q1_Happy` variable is used. The difference in mean scores in `Q1_Happy` between "Living with Ed" and "My Name is Earl" viewers is computed from a bootstrap sample constructed from the original sample, and this process is repeated 10000 times to generate 10000 bootstrapped differences in mean `Q1_Happy` scores. Then, a 95% confidence interval is constructed from the bootstrapped samples to evaluate the difference in mean `Q1_Happy` scores.

Results



The graph above displays differences in the distribution of happiness scores between “My Name is Earl” and “Living with Ed” in the survey sample. Both shows have a median happiness score of 4, but “Living with Ed” appears to have a wider range of score values than “My Name is Earl”.



```
##       name      lower    upper level     method    estimate
## 1 diffmean -0.4017012 0.112142  0.95 percentile -0.06541353
```

The graph above displays the distribution of the bootstrapped differences in mean happiness scores.

Conclusion

The confidence interval for these differences in means demonstrates with 95% confidence that the difference in mean happiness scores between “Living with Ed” and “My Name is Earl” is between -0.3991 and 0.1004. However, since 0 is within this range (a difference of means equal to 0 would mean there is no difference in happiness scores between the two shows), the data do not support the theory that viewers feel different levels of happiness when watching “Living with Ed” and “My Name is Earl”.

Part B

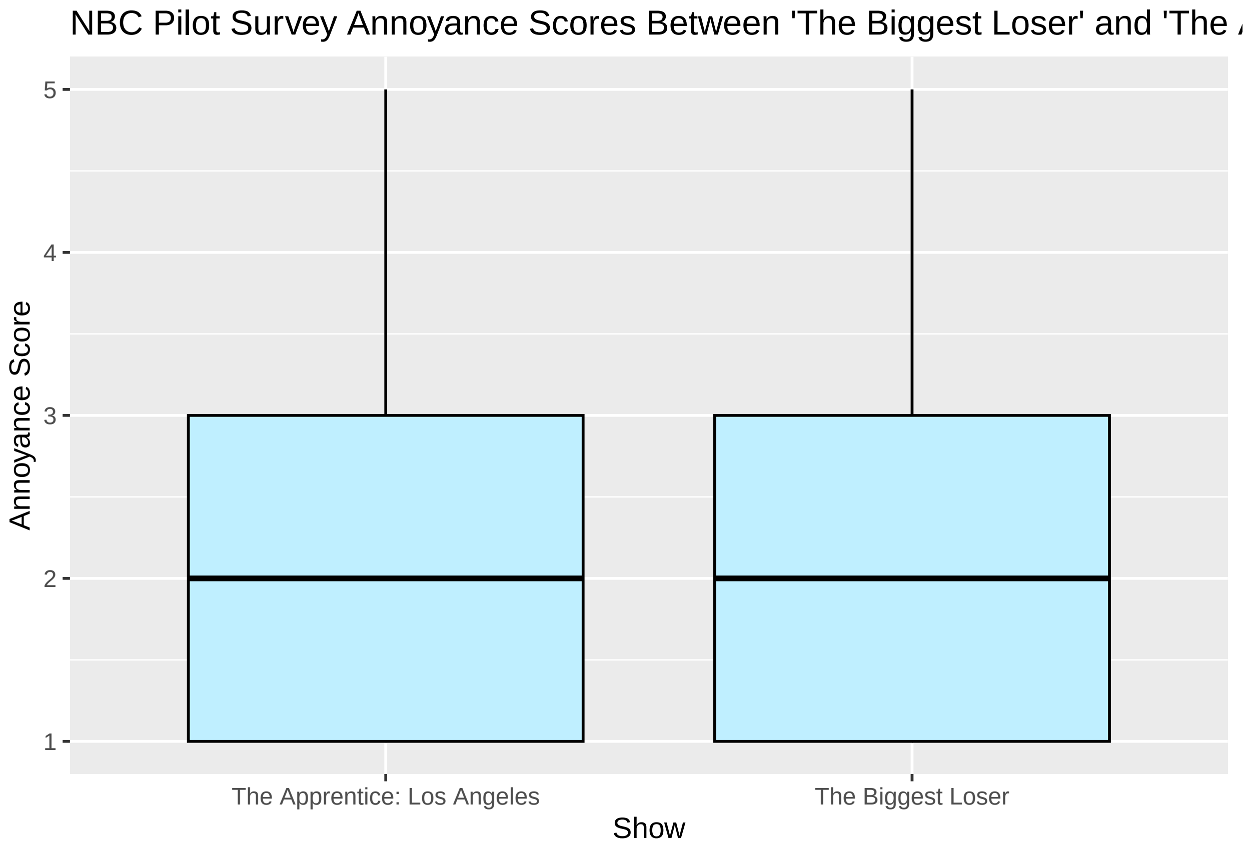
Question

We want to answer: which shows made viewers more annoyed, “The Biggest Loser” or “The Apprentince: Los Angeles”?

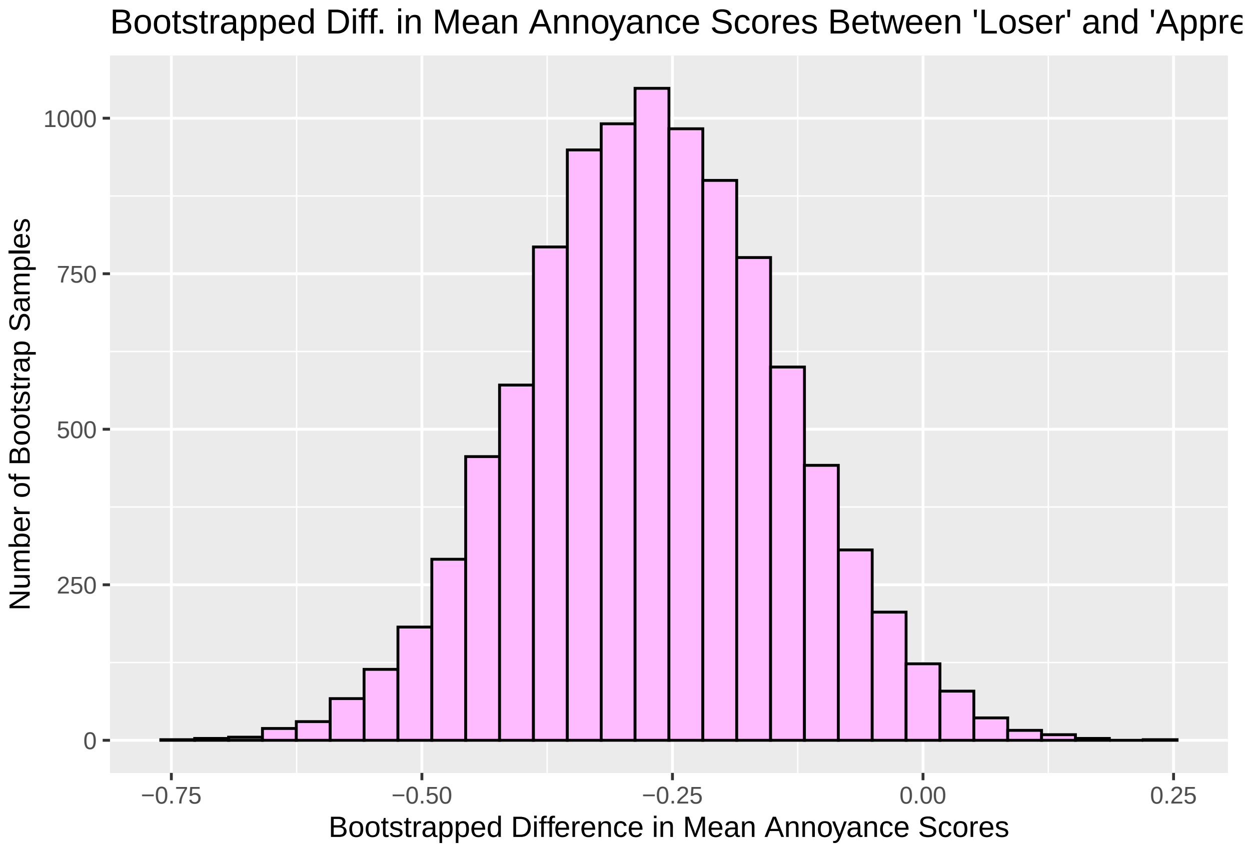
Approach

A dataset is created only including the shows “The Biggest Loser” and “The Apprentince: Los Angeles”, and differences in the values of `Q1_Annoyed` in the original sample are visualized using a factored boxplot. Then, 10000 bootstrapped difference in mean annoyance scores are generated from the original dataset, visualized on a histogram. Finally, a 95% confidence interval is constructed from the bootstrapped difference in mean annoyance scores to examine the strength of evidence that one show causes more annoyance than the other.

Results



Within the NBC sample, “The Apprentince: Los Angeles” and “The Biggest Loser” appear to have identical annoyance score distributions, with median annoyance scores of 2.



```
##       name      lower       upper level     method   estimate
## 1 diffmean -0.5205175 -0.01344397  0.95 percentile -0.2764848
```

The graph above displays the bootstrapped difference in mean annoyance scores between "The Biggest Loser" and "The Apprentice: Los Angeles".

Conclusion

The confidence interval computed demonstrates with 95% confidence that the difference in mean annoyance scores between "The Biggest Loser" and "The Apprentice: Los Angeles" lies between -0.5231 and -0.0218. Since 0 is not within this interval, the data do support the idea that viewers had different levels of annoyance between "The Biggest Loser" and "The Apprentice: Los Angeles".

Part C

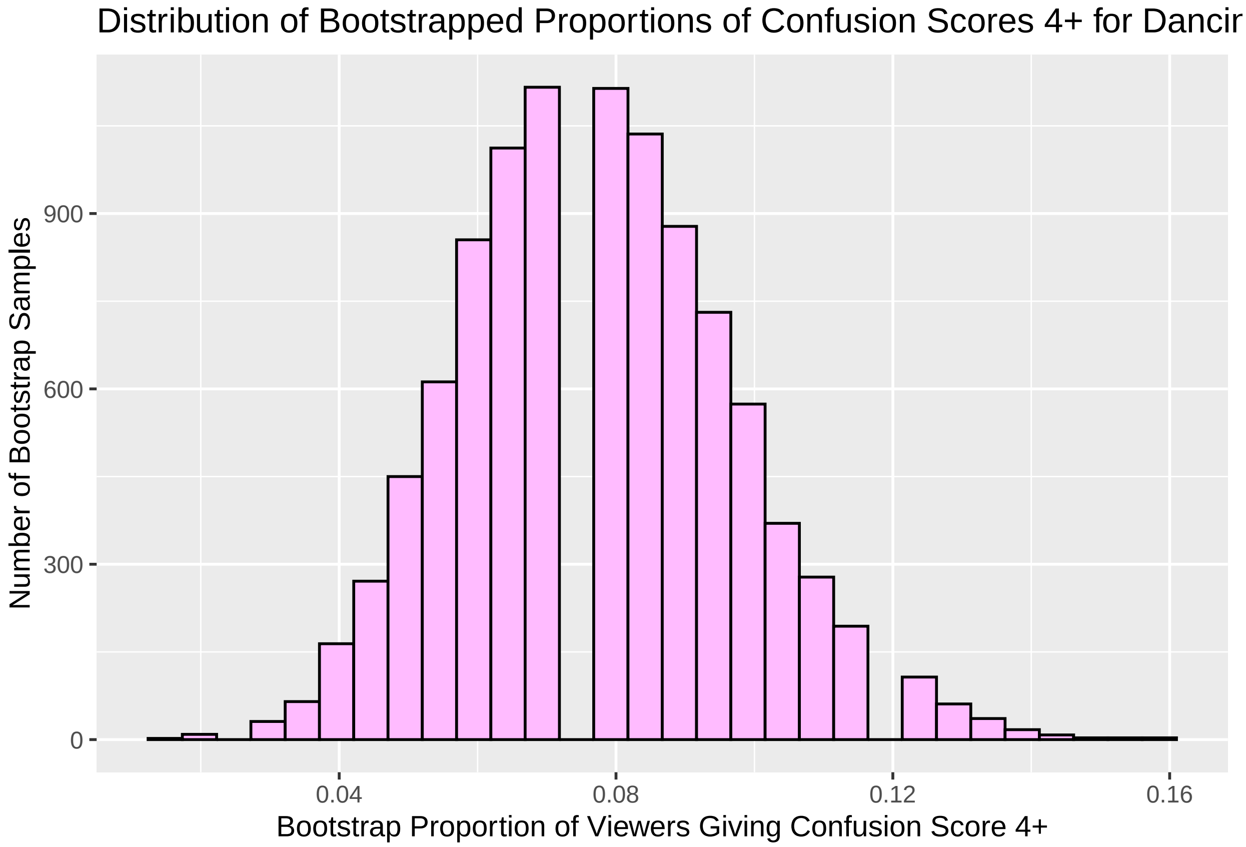
Question

What is the proportion of American TV watchers expected to have given a score of 4 or above to the NBC's Q2_Confusing pilot survey question for the show "Dancing with the Stars"?

Approach

A dataset is created from the original survey only containing observations with the show "Dancing with the Stars", and a new binary variable is created that assigns a TRUE or FALSE value depending on whether or not a viewer gave a confusion score of 4 or higher. 10000 bootstrap samples are generated, and for each the proportion of viewers who gave “Dancing with the Stars” a confusion score of 4 or higher is calculated and stored as entries into a dataframe. Then, a 95% confidence interval is constructed from the bootstrapped proportions.

Results



```
##        name      lower     upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.1160221  0.95 percentile 0.08839779
```

Above is the distribution of the bootstrapped proportions of viewers giving “Dancing with the Stars” a confusion score of 4 or greater. It is centered at around 0.07-0.08.

Conclusion

The confidence interval generated determines with 95% confidence that the proportion of American TV watchers expected to have given a score of 4 or above to the NBC’s Q2_Confusing pilot survey question for the show “Dancing with the Stars” lies between 0.03887 and 0.11602.

Problem 4

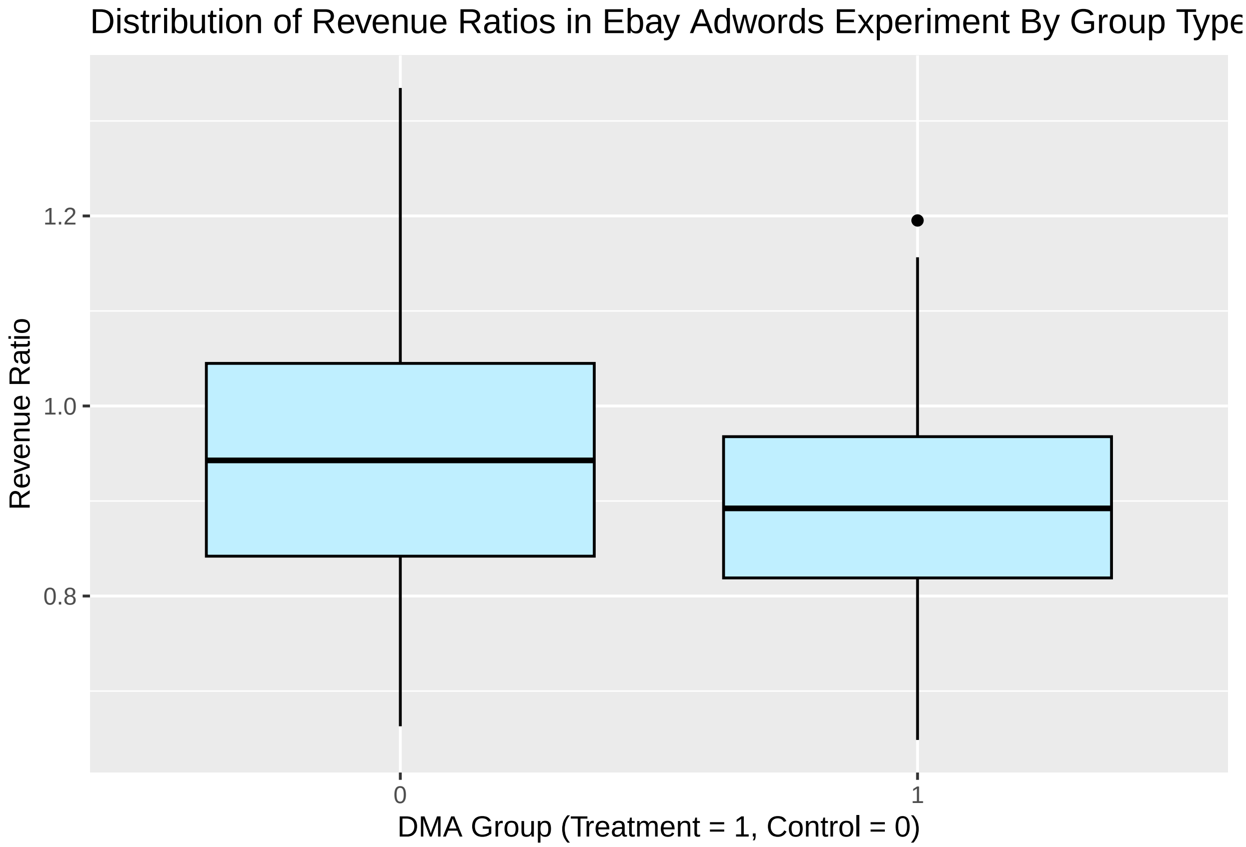
Question

Does Ebay’s paid advertising on Google AdWords increase revenue for Ebay?

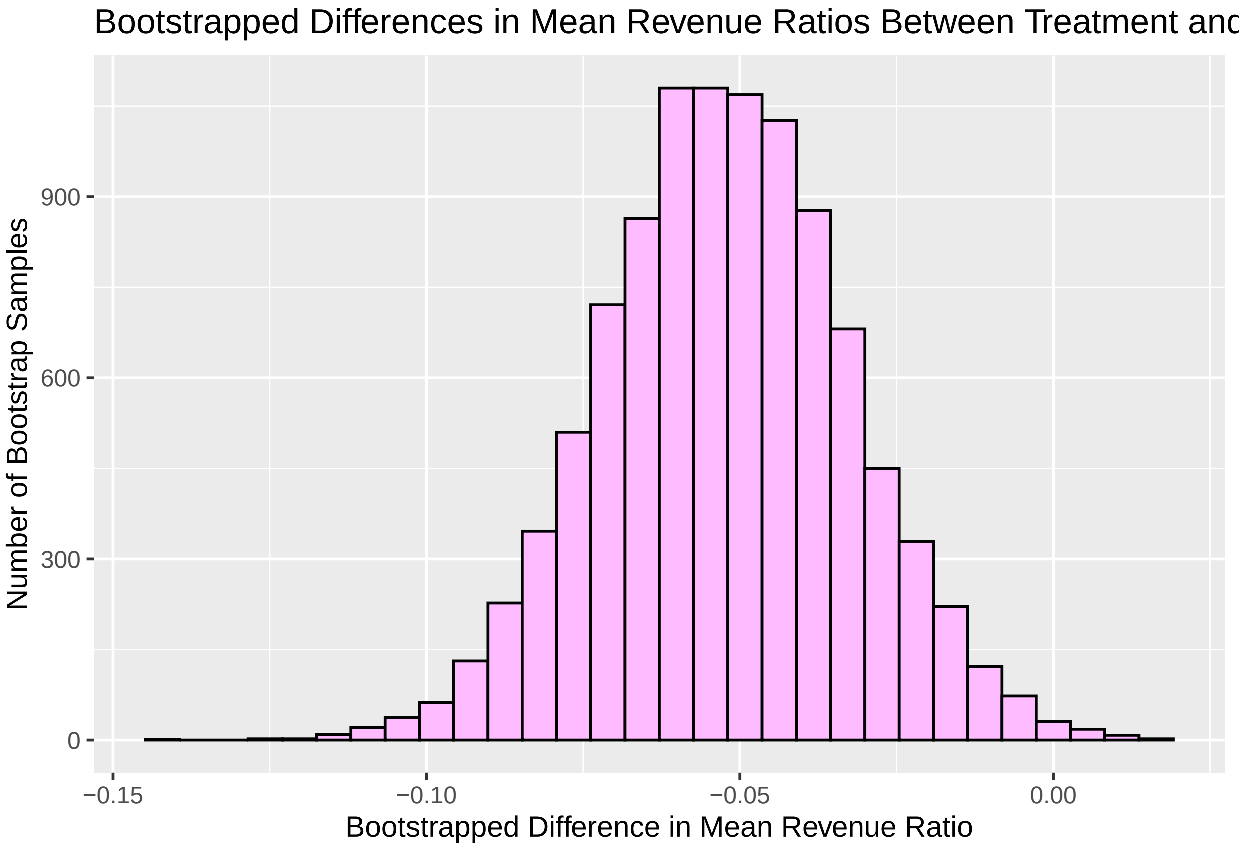
Approach

To determine whether revenue changed before to after the experiment, a revenue ratio variable is created; that is, the ratio of revenue 30 days after the experiment to revenue 30 days before the experiment. 10000 bootstrap samples are taken, with the difference in the mean revenue ratio between treatment and control DMAs calculated for each sample. The distribution of these differences in means is visualized in a histogram, and a 95% confidence interval is constructed for the difference in mean revenue ratio.

Results



The factored boxplot above shows the distribution of revenue ratios across DMAs in Ebay's Adwords experiment by the DMA's experimental group, with treatment groups that experienced a paid ad pause designated by 1, and control groups that had no paid ad pause indicated by 0. DMAs in the treatment group appeared to have lower revenue ratios than control DMAs, indicated that the former saw relatively lower revenue after the experiment.



```
##       name       lower       upper level     method   estimate
## 1 diffmean -0.09059093 -0.01357998  0.95 percentile -0.0559797
```

The graph above displays the distribution of 10000 bootstrapped differences in mean revenue ratios.

Conclusion

The confidence interval computed determines with 95% confidence that the difference in the mean revenue ratio between DMAs that had paid ads on Google AdWOrds paused for a month and DMAs that continued running paid ads on Google Adwords lies between -0.09115 and -0.01396. In other terms, it appears that DMAs that paused paid Google ads saw generally lower revenue ratios than DMAs that did not pause paid Google ads. Therefore, the data provide support for the theory that Ebay's paid advertising on Google AdWords increases revenue for Ebay.