SDS 315 Homework2

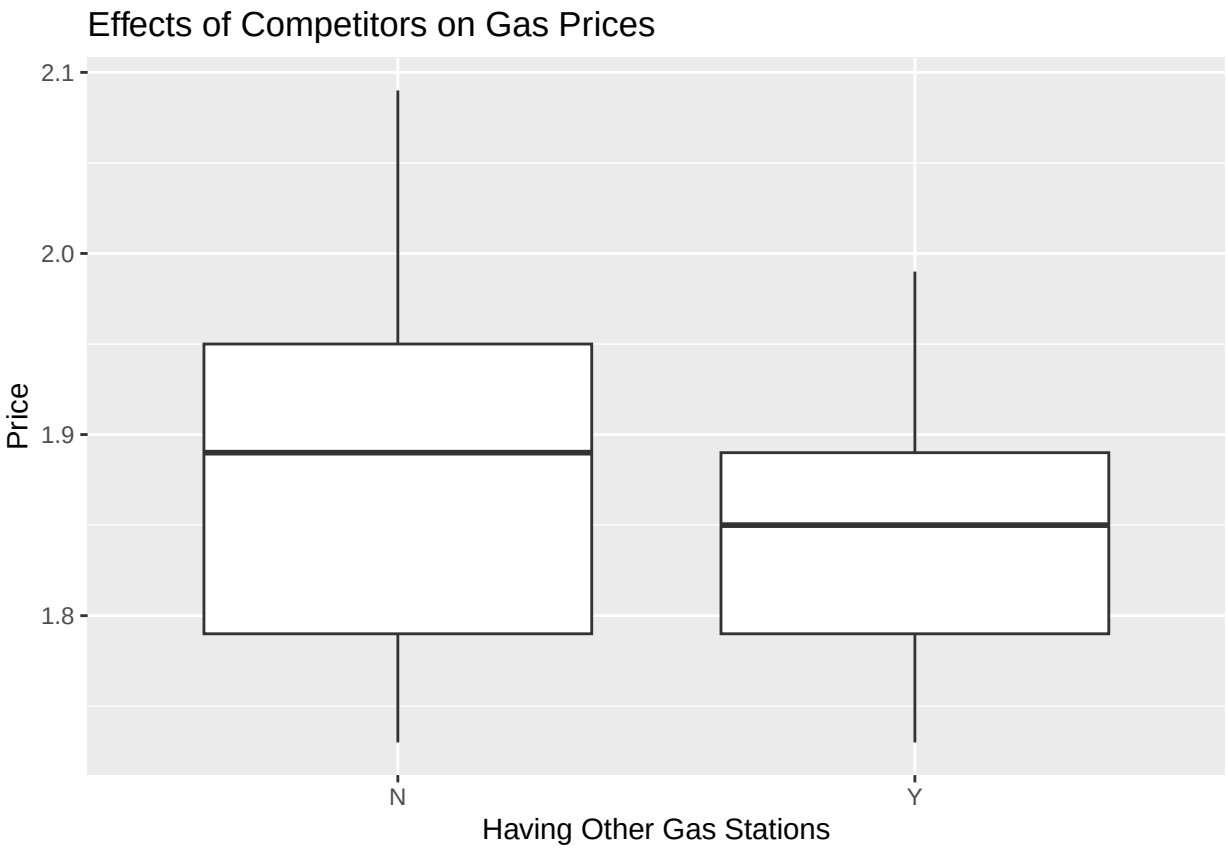
Daniel Wu (EID: djw3627)

2025-02-06

To access my GitHub repository, click here: https://github.com/DanielWu3627/SDS315/tree/main. Please check the file named **Homework3.Rmd**.

Problem 1

Theory A



This figure shows the distribution of prices, which depends on whether there are any other gas stations in sight (Y) or not (N). We can see the median gas price is higher when there is no other gas stations in sight.

```
##    diffmean
## -0.02348235
```

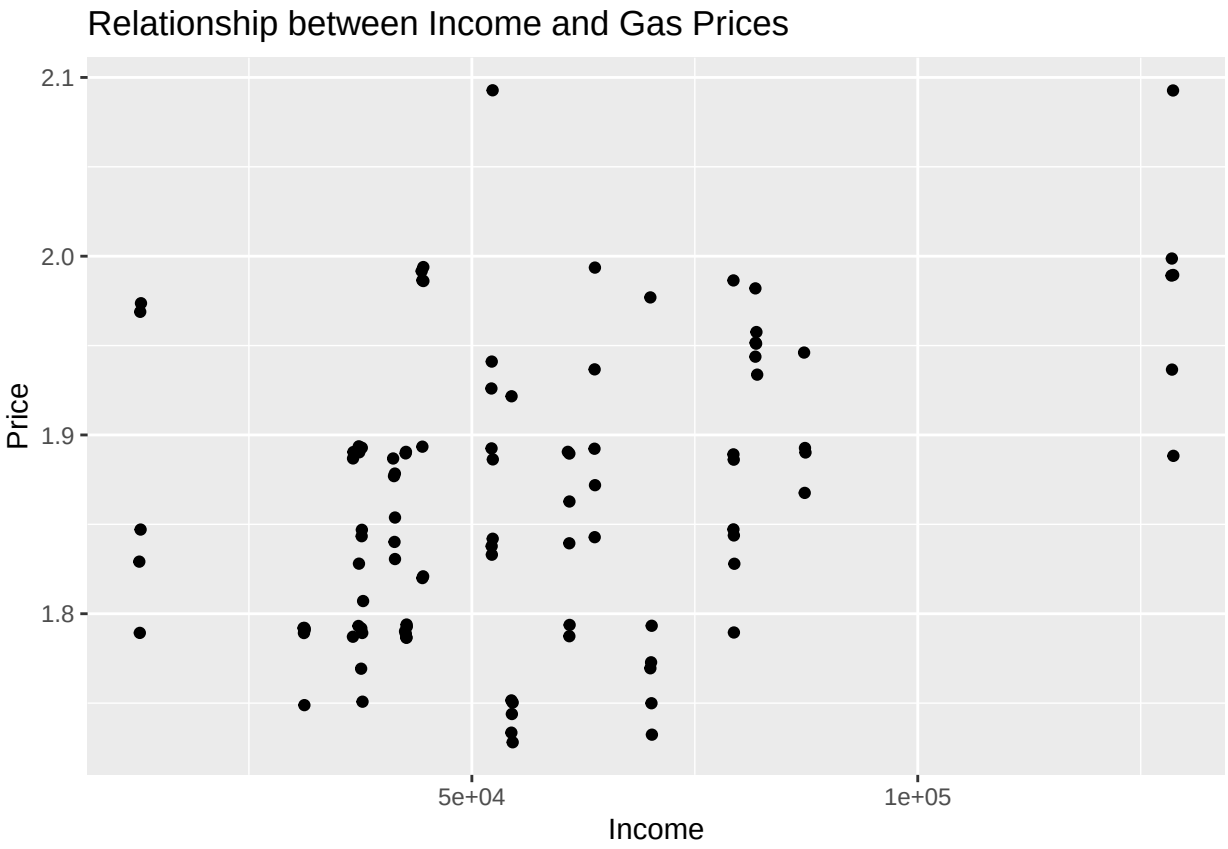
```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05496178 0.008314663  0.95 percentile -0.02348235
```

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

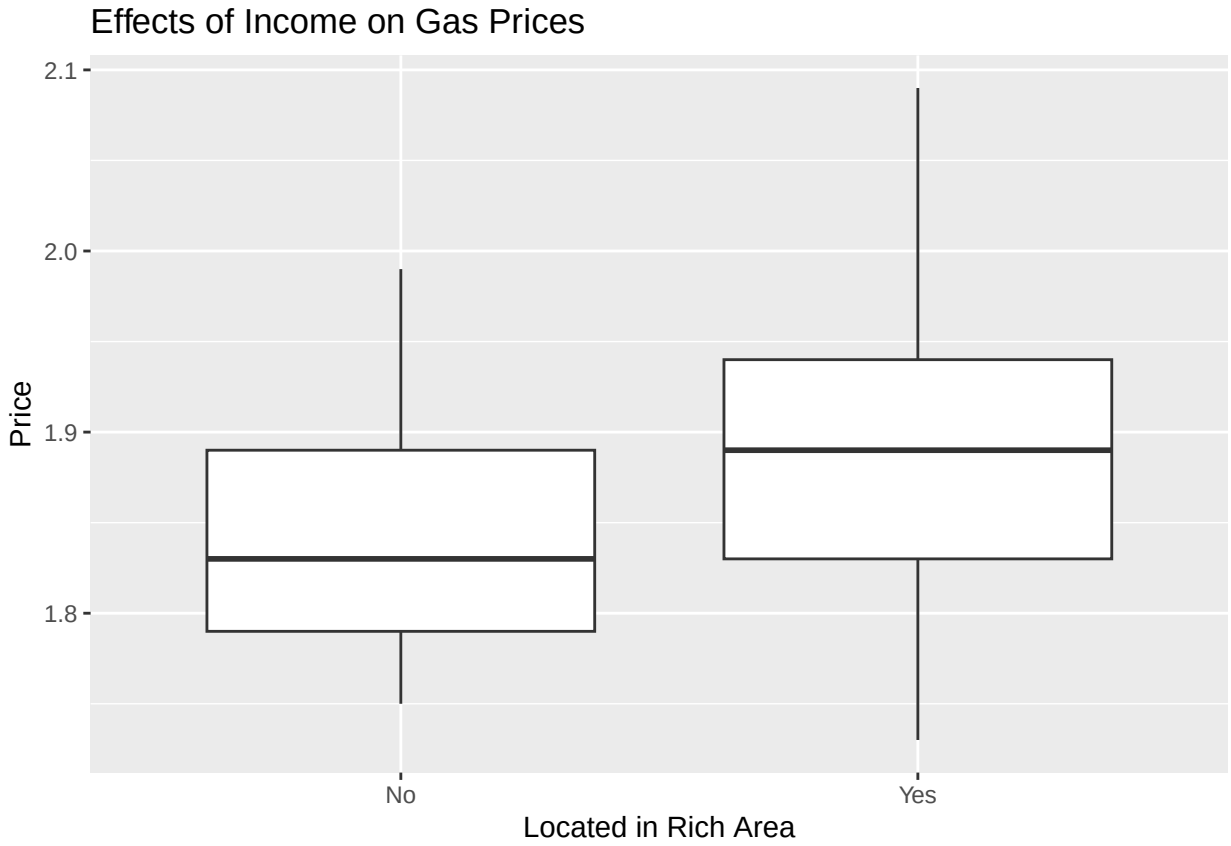
Evidence: With a 95% confidence interval, the difference in price between gas stations with and without direct competitors is somewhere between -0.056 and 0.008. Therefore, the difference is not statistically significant at the 0.05 level because the 95% confidence interval for that difference contains zero.

Conclusion: The theory is not supported by the data. The difference in gas prices is not significant between gas stations with and without direct competitors.

Theory B



This scatterplot shows the relationship between income and gas prices. It seems that gas prices are positively correlated to Income. To make comparison easier, I used the median income of all households in the sample as the cutoff. If the area with the household income that is higher than this cutoff, it is a rich area.



This figure shows the distribution of prices, which depends on whether the gas station is in a rich area. We can see the median gas price is higher when the gas station is located in a rich area.

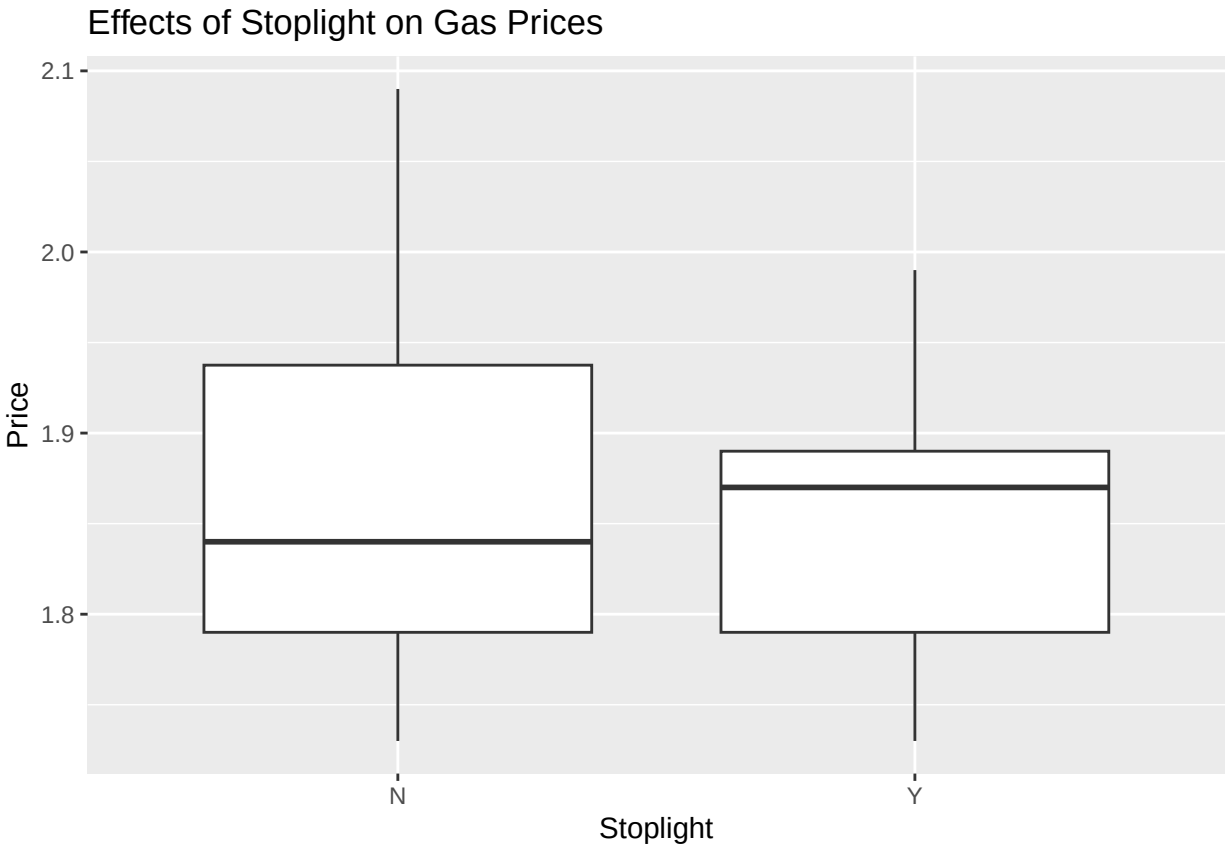
```
##       name      lower      upper level     method   estimate
## 1 diffmean 0.00322009 0.06511677  0.95 percentile 0.03462569
```

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

Evidence: With a 95% confidence interval, the difference in price between gas stations located in high and low income areas (lower than the median household income of $52,306) is somewhere between 0.004 and 0.065. Therefore, the difference is statistically significant at the 0.05 level because the 95% confidence interval for that difference does not contain zero.

Conclusion: The theory is supported by the data. The gas prices in gas stations located in high income areas is significantly higher than that in low income areas.

Theory C



This figure shows the distribution of prices, which depends on whether there is a stop light (Y) or not (N) in front of a gas station. We can see the median gas price is higher when there is a stoplight in front of a gas station.

```
##     diffmean
## -0.003299916
```

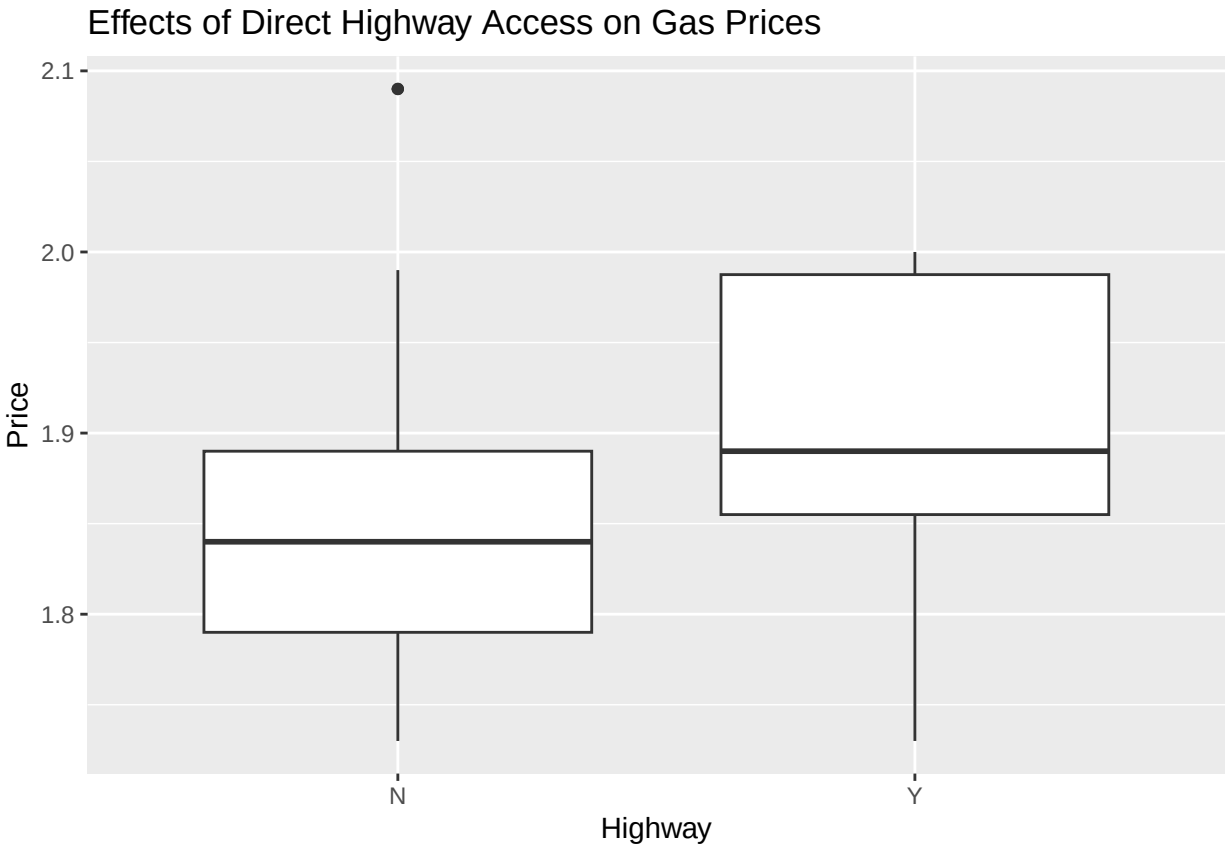
```
##       name       lower      upper level     method     estimate
## 1 diffmean -0.03864805 0.03015543  0.95 percentile -0.003299916
```

Claim: Gas stations at stoplights charge more.

Evidence: With a 95% confidence interval, the difference in price between gas stations with and without a stoplight in sight is somewhere between -0.039 and 0.032. Therefore, the difference is not statistically significant at the 0.05 level because the 95% confidence interval for that difference contains zero.

Conclusion: The theory is not supported by the data. The difference in gas prices is not significant between gas stations with and without direct competitors.

Theory D



This figure shows the distribution of prices, which depends on whether the gas station is accessible from highway (Y) or not (N). We can see the median gas price is higher when the gas station is accessible from a highway.

```
##  diffmean
## 0.0456962
```

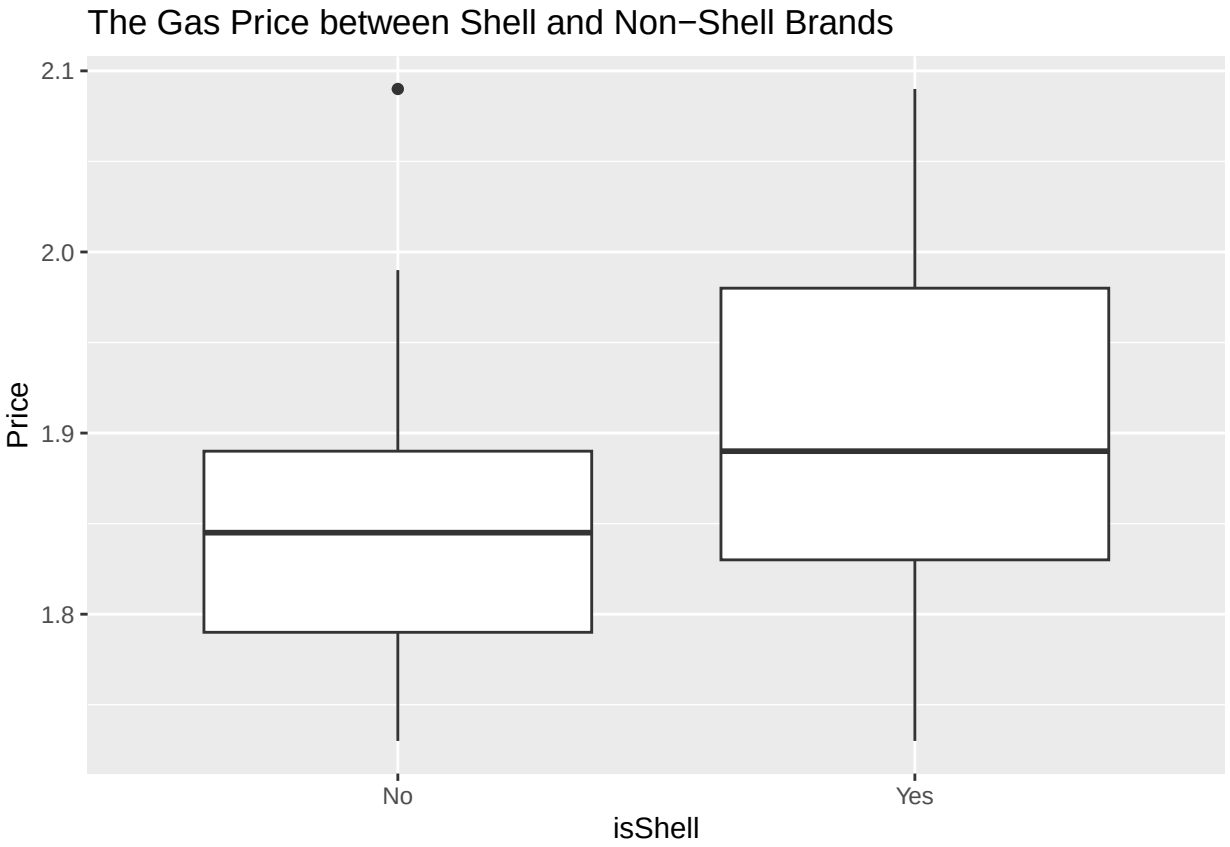
```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008906584 0.08090525  0.95 percentile 0.0456962
```

Claim: Gas stations with direct highway access charge more.

Evidence: With a 95% confidence interval, the difference in price between gas stations with and without direct access to highways is somewhere between 0.009 and 0.081. Therefore, the difference is statistically significant at the 0.05 level because the 95% confidence interval for that difference does not contain zero.

Conclusion: Therefore, the theory is supported by the data. The gas prices are higher at stations with direct highway access than those without direct highway access.

Theory E



This figure shows the distribution of prices, which depends on whether the brand is shell. We can see the median gas price is higher when the brand is Shell.

```
##   diffmean
## 0.02740421
```

```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008906584 0.08090525  0.95 percentile 0.0456962
```

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

Evidence: With a 95% confidence interval, the difference in price between Shell and other brands is somewhere between 0.009 and 0.081. Therefore, the difference is statistically significant at the 0.05 level because the 95% confidence interval for that difference does not contain zero.

Conclusion: Therefore, the theory is supported by the data. The gas prices are significantly higher at Shell stations than other brands.

Problem 2

Part A

```
##   name    lower    upper level     method estimate
## 1 mean 26254.05 31798.18  0.95 percentile 27481.69
```

With 95% confidence, the average mileage of 2011 S-Class 63 AMGs that were hitting the used car market is between 26250 and 31730 miles.

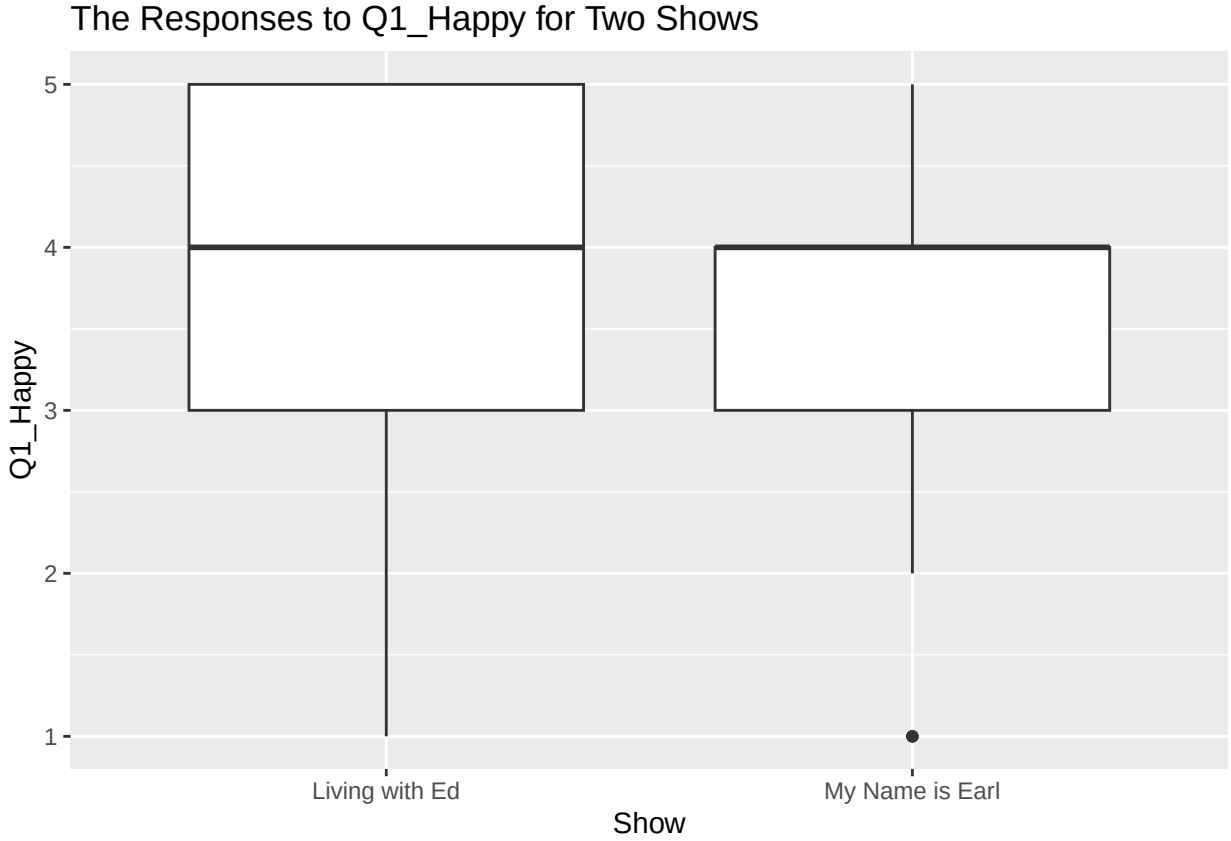
Part B

```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4167532 0.4527518  0.95 percentile 0.4347525
```

With 95% confidence, the proportion of all 2014 S-Class 550s that were painted black is between 0.417 and 0.453 (41.7% and 45.3%).

Problem 3

Part A



This figure shows the distribution of Q1_Happy responses, which depends on the show (Living with Ed or My Name is Earl). We can see the median for both shows are about the same.

```
##   diffmean
## -0.1490515
```

```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.4038387 0.1008292  0.95 percentile -0.1490515
```

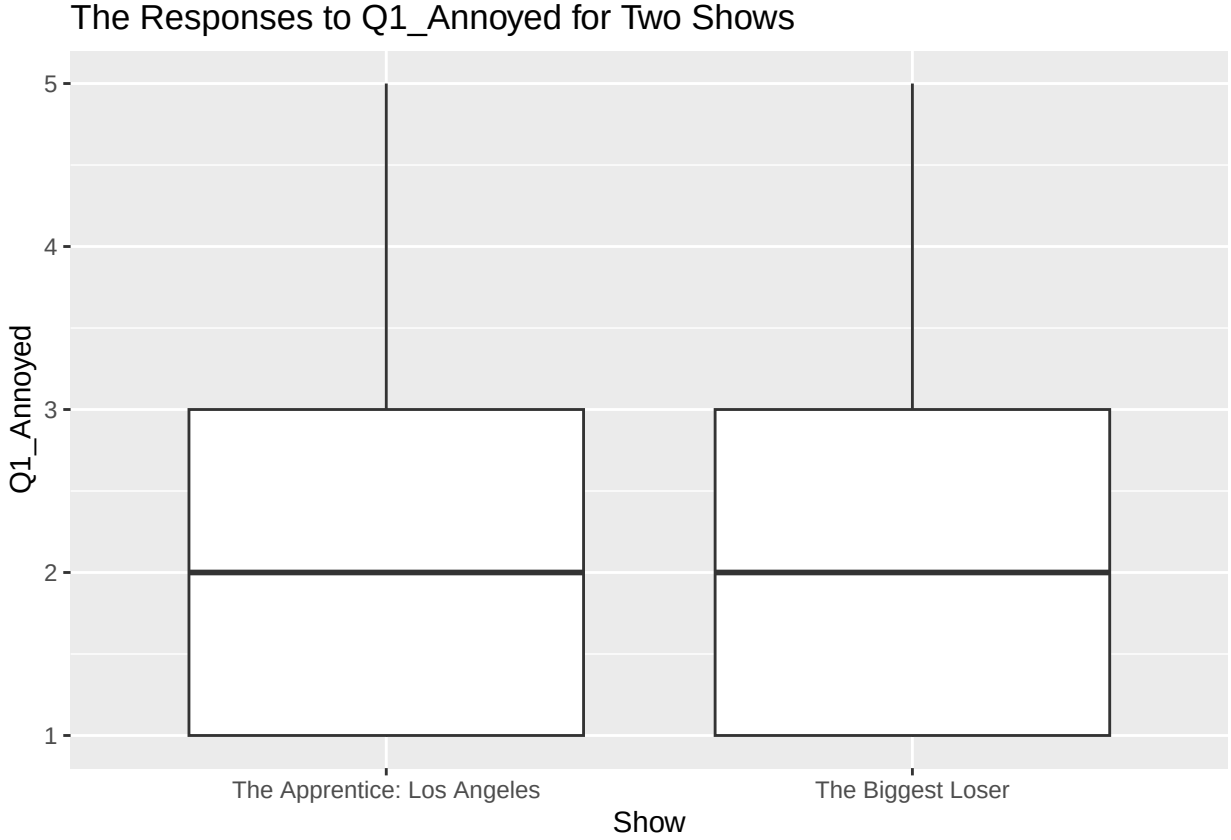
Question: Based on this sample of respondents, which show makes people happier; *Living with Ed* or *My Name is Earl*?

Approach: I filtered the original dataset to include only rows with shows that are "Living with Ed" or "My Name is Earl." I constructed a 95% confidence interval for the difference in mean viewer response to the Q1_Happy question for these two shows.

Results: The confidence interval for the difference in means of viewer responses in Q1_Happy is between -0.399 and 0.103.

Conclusion: Since the confidence interval includes 0, the difference between response for the two shows are not statistically significant at the 0.05 significance level. Thus, the results do not provide strong evidence to conclude that one show generates a consistently higher mean Q1_Happy response than the other.

Part B



This figure shows the distribution of Q1_Annoyed responses, which depends on the show (The Apprentice: Los Angeles or The Biggest Loser). We can see the median Q1_Annoyed for both shows are about the same.

```
##  diffmean
## -0.270997
```

```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5199084 -0.02051757  0.95 percentile -0.270997
```

Question: Based on this sample, which show makes people feel more annoyed: *The Biggest Loser* or *The Apprentice: Los Angeles*?

Approach: I filtered the original dataset to include only rows with shows that are "The Biggest Loser" or "The Apprentice: Los Angeles." I constructed a 95% confidence interval for the difference in mean viewer response to the Q1_Annoyed question for these two shows.

Results: Using *The Apprentice: Los Angeles* as reference, the 95% confidence interval for the difference in mean viewer response to Q1_Annoyed for the two shows is between -0.526 and -0.020. This suggests that with 95% confidence, the mean of viewer response to the Q1_Annoyed question for *The Apprentice: Los Angeles* is between 0.020 and 0.526 higher than that for *The Biggest Loser*.

Conclusion: Since 0 is not within the confidence interval, there is statistically significant evidence of a difference in the mean of Q1_Annoyed responses between the two shows. Additionally, due to the consistently negative confidence interval, there is evidence that views of *The Apprentice: Los Angeles* tend to make people feel more annoyed (higher Q1_Annoyed response) compared to viewers of *The Biggest Loser*.

Part C

```
##        name      lower     upper level     method   estimate
## 1 prop_TRUE 0.03867403 0.1160221  0.95 percentile 0.07734807
```

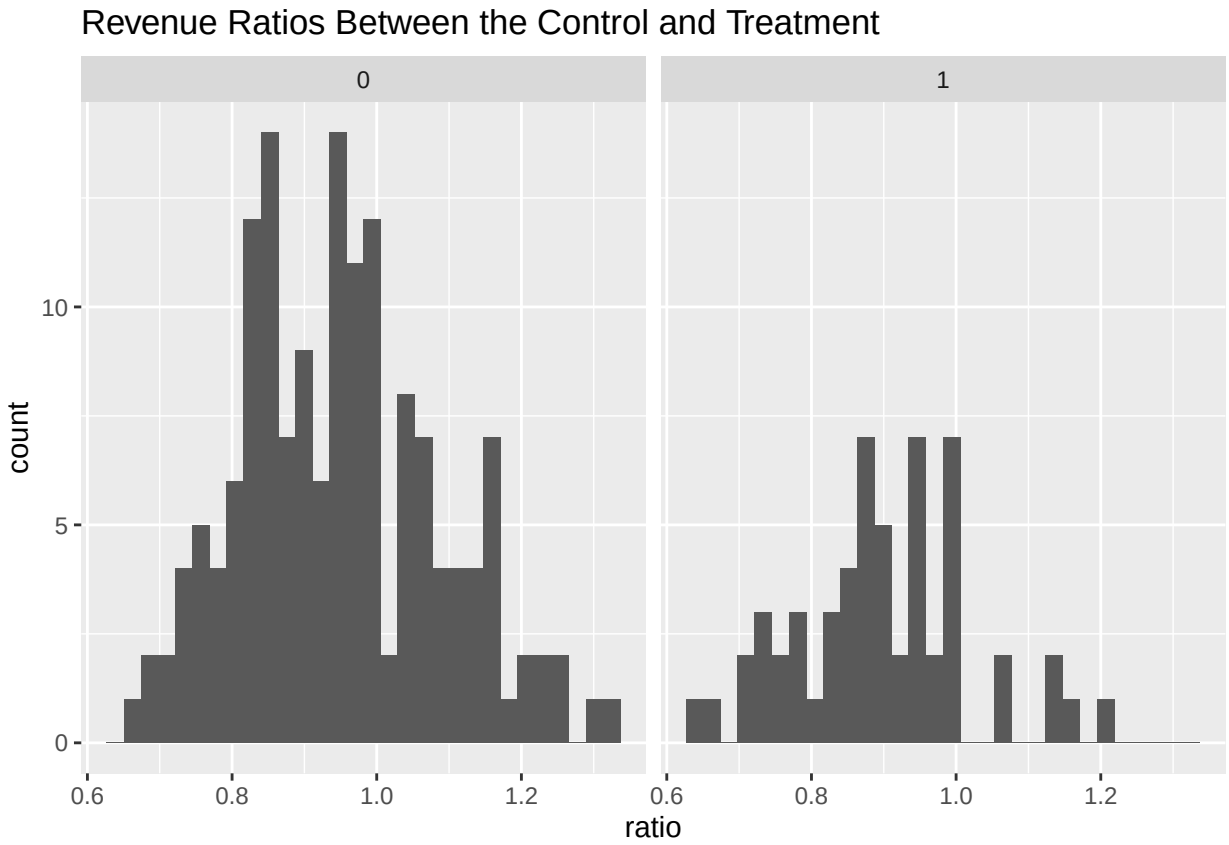
Question: Based on this sample of respondents, what proportion of American TV watchers would find *Dancing with the Stars* confusing (a response of 4 or greater to the Q2_Confusing question)?

Approach: I filtered the original dataset to include only rows with shows that are "Dancing with the Stars." I constructed a 95% confidence interval for the difference in mean viewer response to the Q2_Confusing question for the show.

Results: The 95% confidence interval for the proportion of American TV watchers who we expect to give a response of 4 or greater to the "Q2_Confusing" question is approximately between 0.039 and 0.116 (3.90% and 11.6%).

Conclusion: We are 95% confident that the true proportion of American TV watchers who would rate "Dancing with the Stars" as confusing (rating 4 or greater) lies between 3.87% and 12.89%. Therefore, the proportion is statistically significant at the 0.05 significance level.

Problem 4



The histogram shows the distribution of revenue ratio (ratio of the revenue after to the revenue before the experiment) between the control (0) and treatment group (1).

```
##       name      lower       upper level     method    estimate
## 1 diffmean -0.0906773 -0.01313271  0.95 percentile -0.05228145
```

Question: Based on this sample of respondents, is the revenue ratio significantly different between the treatment and control DMAs? If yes, does the paid search advertising on Google create extra revenue for

Ebay?

Approach: I made a new variable that is the ratio of the revenue after to the revenue before the experiment. I constructed a 95% confidence interval for the difference in revenue ratios between treatment and control groups.

Results: The 95% confidence interval for the difference in ratios between the treatment and control group is between -0.091 and -0.130.

Conclusion: We are 95% confident that the true difference in the ratio of revenue after and before the experiment is between -0.091 and -0.130. Since the interval does not contain zero, the difference in revenue ratio between the treatment and control group is statistically significant. Since the control group is the reference and the estimated difference is negative, we can conclude that the revenue ratio for the treatment group is significantly lower than the control group, suggesting that paid search advertising may create extra revenue for Ebay.