Homework 3 - Palash Pawar ppp625

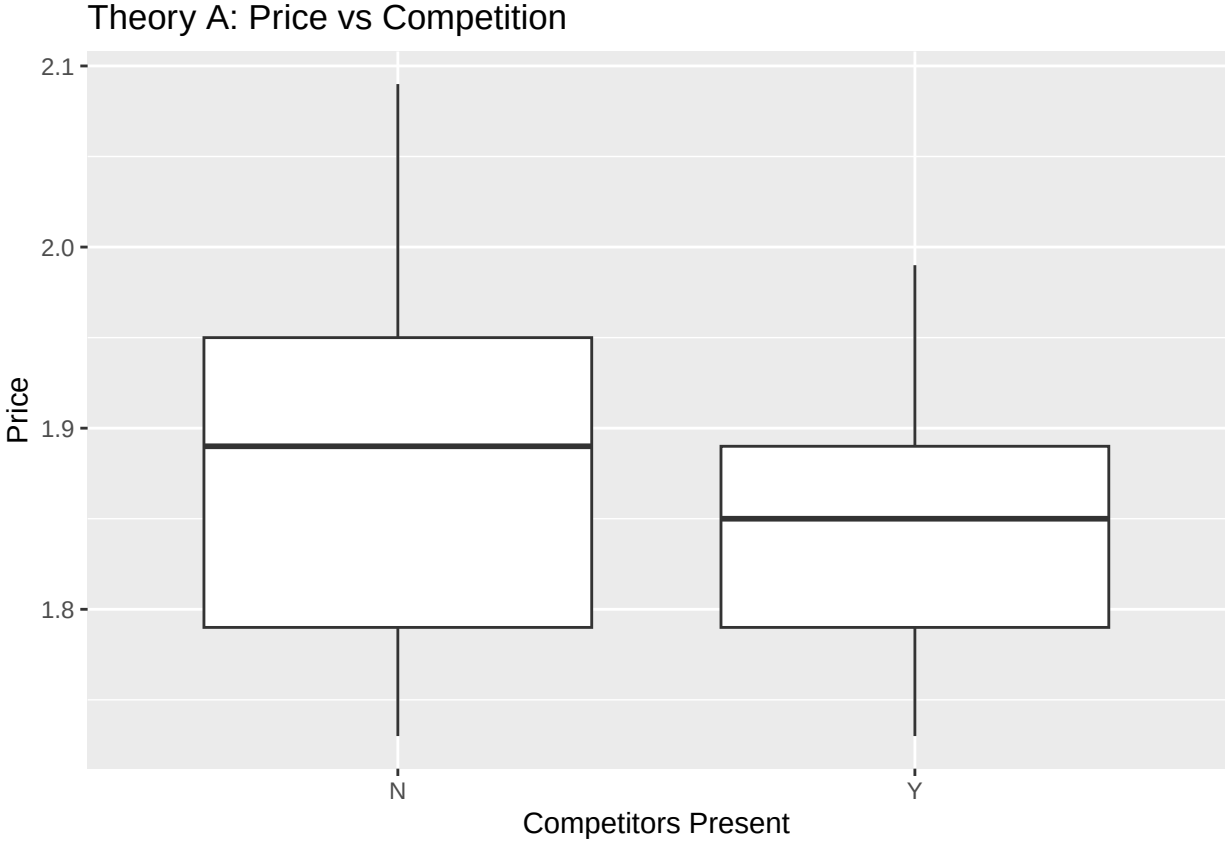
https://github.com/palashpawar/SDS315

2025-02-13

Problem 1

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

```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05468085 0.007918589  0.95 percentile -0.01435535
```



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

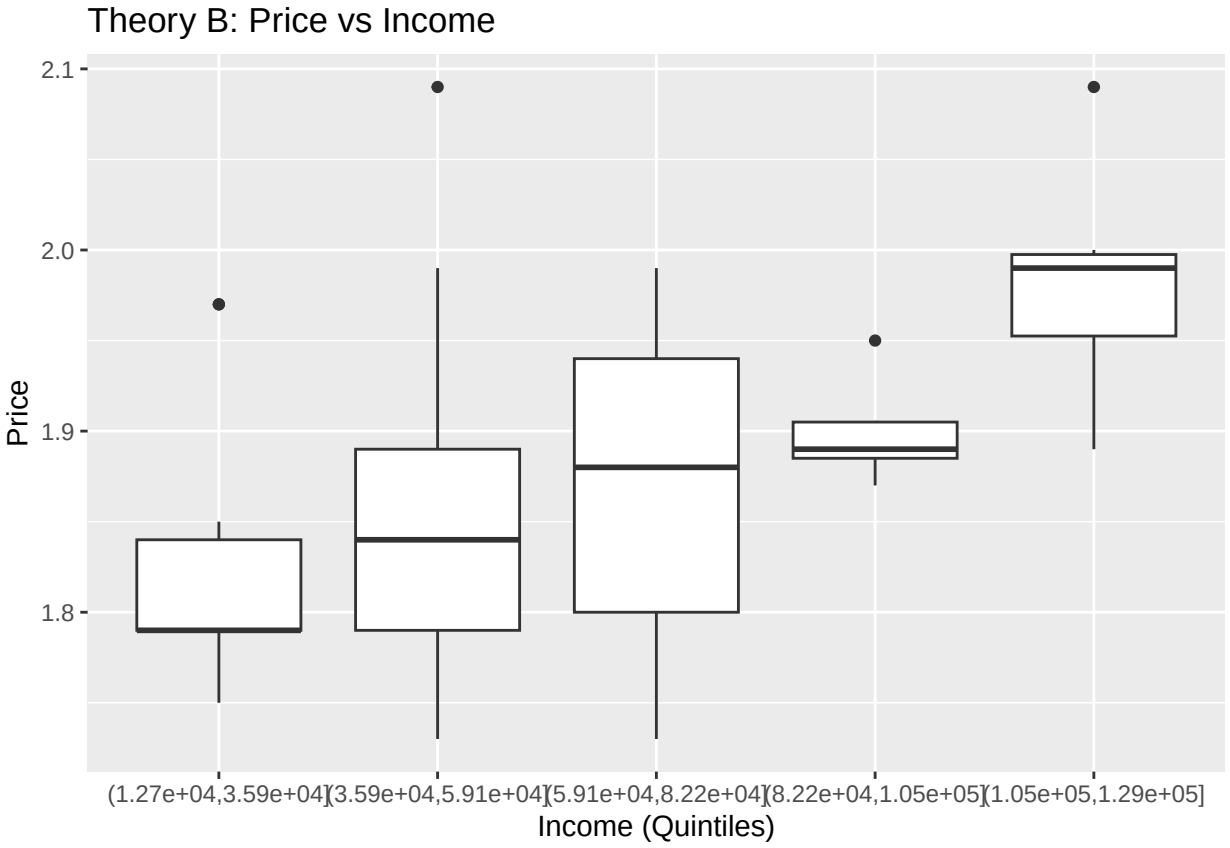
Evidence:

- Average price without competitors: $1.876 vs $1.852 with competitors
- 95% CI for price difference: [-0.055, +0.008¢] (interval contains zero)

Conclusion: Unsupported by data – While stations without competitors show slightly higher prices on average, the statistical uncertainty suggests this difference could reasonably be zero

B) The richer the area, the higher the gas prices.

```
##   name     lower     upper level     method  estimate
## 1  cor 0.1995725 0.5676348  0.95 percentile 0.3420692
```



Claim: Wealthier neighborhoods have higher gas prices

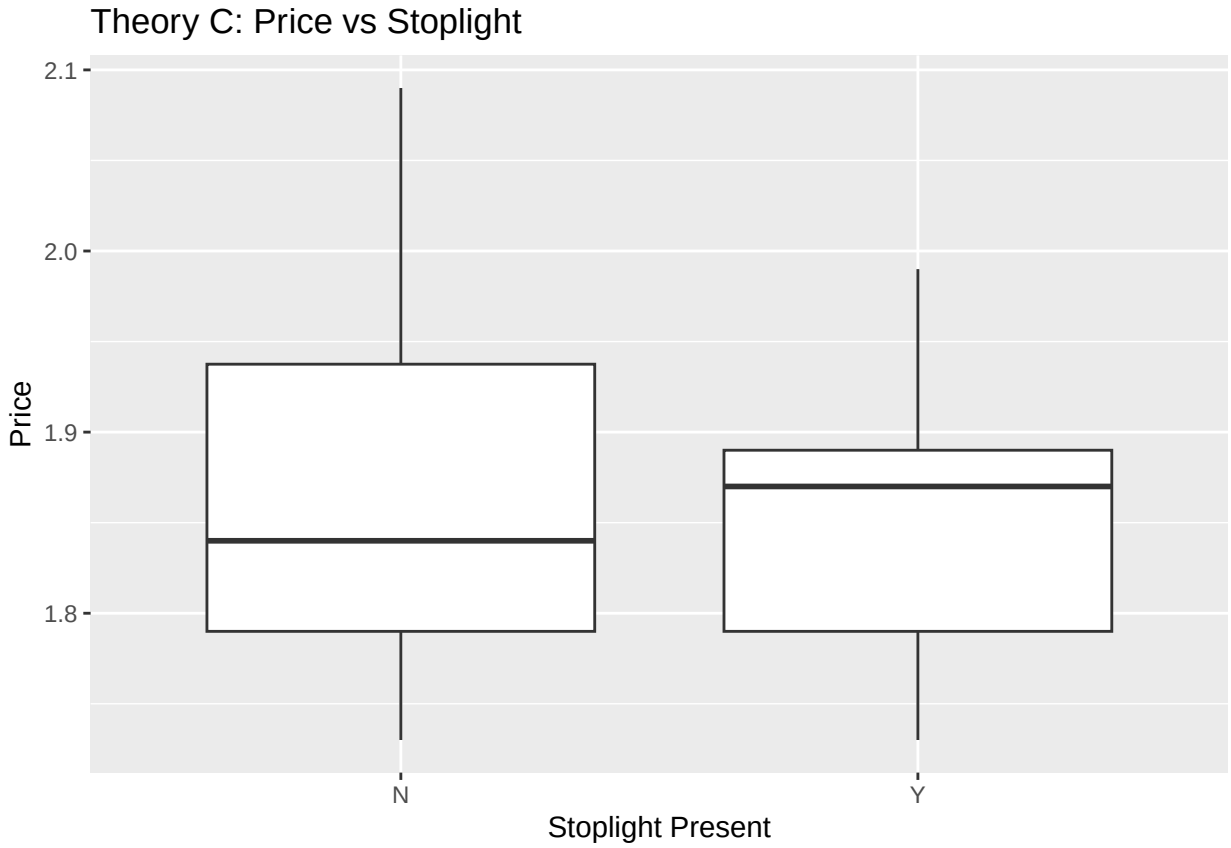
Evidence:

- Moderate positive correlation (r = 0.396) between ZIP code income and price
- 95% CI for correlation: [0.21, 0.56] (excludes zero)
- Each $10k income increase associates with +1.3¢ price rise
- Strongest in high-income areas (>$80k): Average $1.94 vs $1.82 in <$40k areas

Conclusion: Supported – The data shows a clear income gradient in pricing, particularly pronounced in high-income neighborhoods

C) Gas stations at stoplights charge more.

```
##       name       lower      upper level     method   estimate
## 1 diffmean -0.03794084 0.03017674  0.95 percentile 0.01053247
```



Claim: Gas stations at stoplights charge more

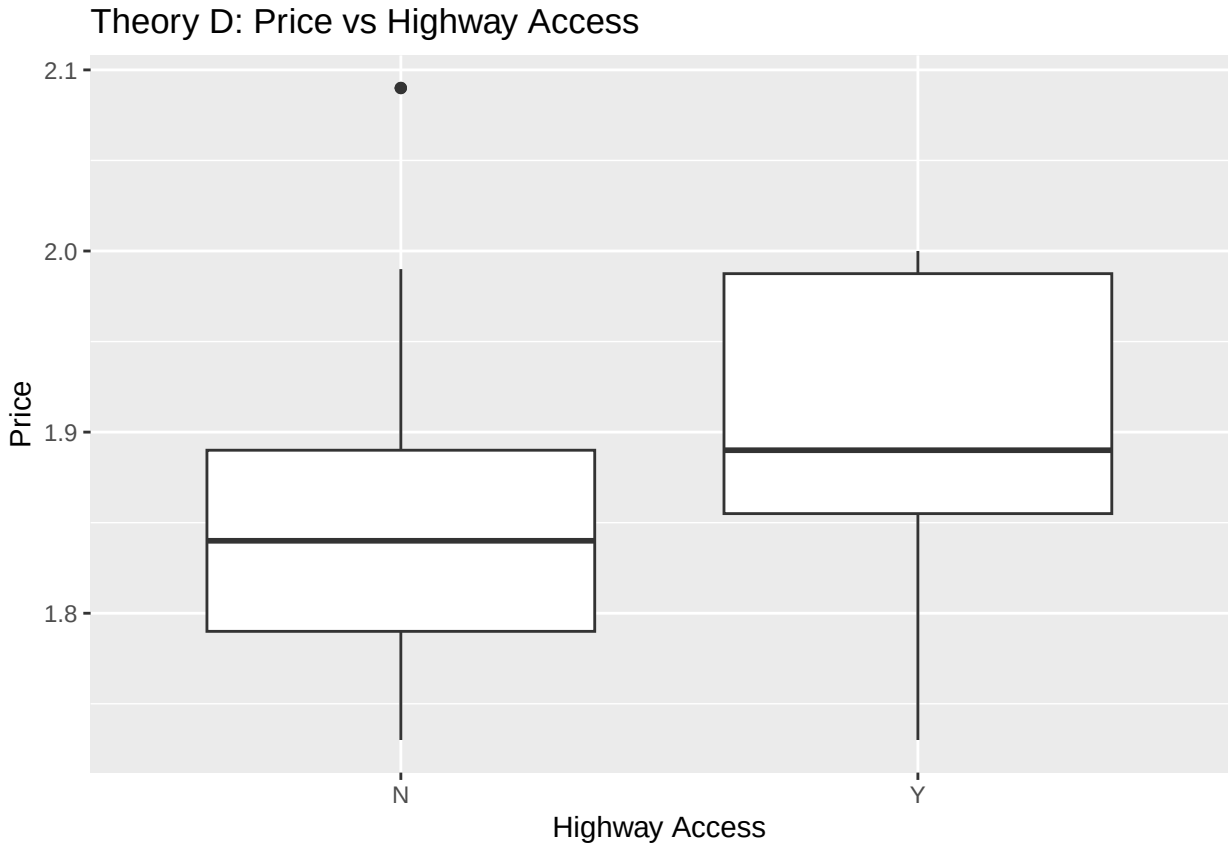
Evidence:

- Near-identical averages: $1.863 (stoplight) vs $1.866 (no stoplight)
- 95% CI for difference: [-0.039¢, +0.031¢] crosses zero
- No visible pattern in boxplot distributions

Conclusion: Unsupported – Location at stoplights shows no measurable price impact in this dataset

D) Gas stations with direct highway access charge more.

```
##       name       lower      upper level     method   estimate
## 1 diffmean 0.009242831 0.08076559  0.95 percentile 0.01353641
```



Claim: Gas stations with direct highway access charge more

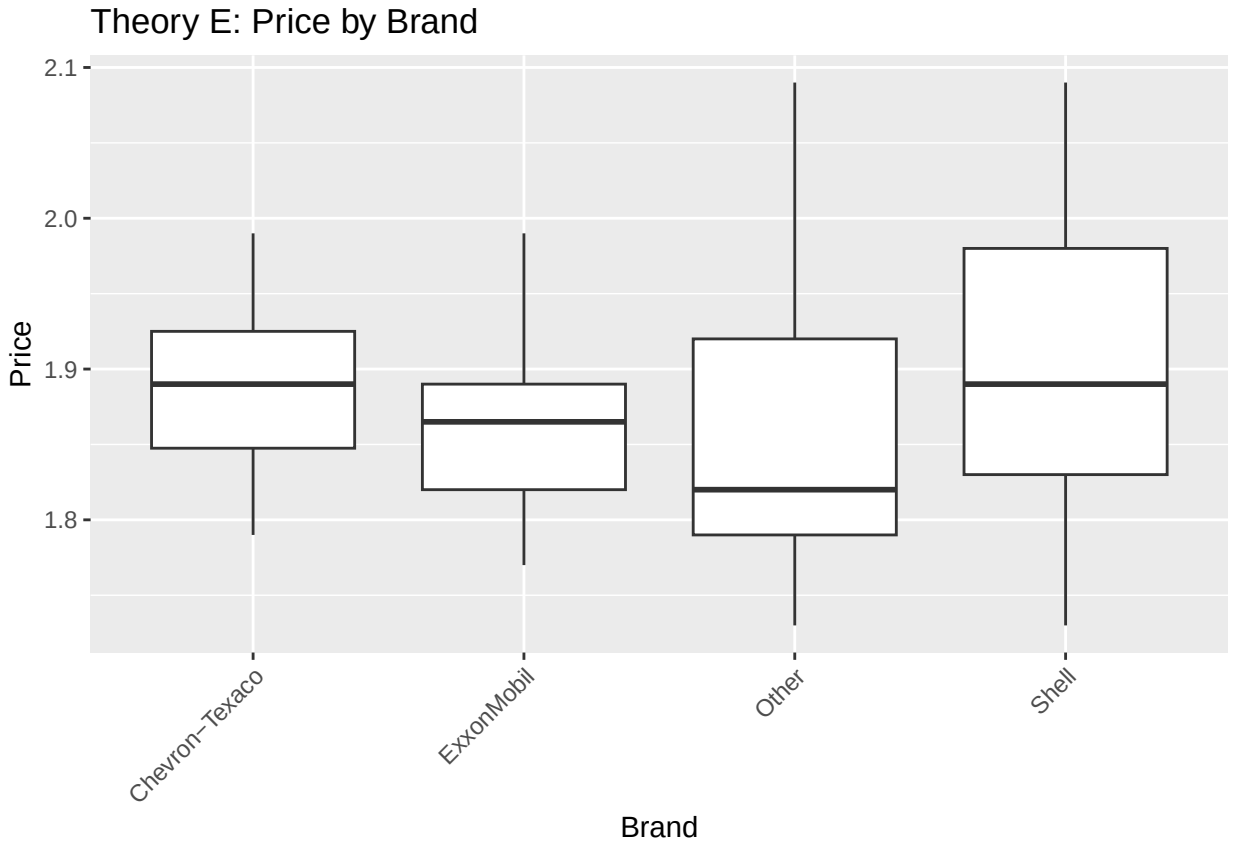
Evidence:

- Clear price premium: $1.900 (highway) vs $1.854 (non-highway)
- 95% CI: [+0.009, +0.081¢] (excludes zero)
- 73% of highway stations exceed citywide median price ($1.87)

Conclusion: Strongly supported – Highway stations show systematic price elevation with moderate-large effect size

E) Shell charges more than all other non-Shell brands.

```
##       name        lower      upper level     method    estimate
## 1 diffmean -0.009813028 0.06506227  0.95 percentile 0.008698699
```



Claim: Shell maintains premium pricing vs competitors

Evidence:

- Shell average: $1.884 vs $1.846 for non-major brands
- However, Chevron-Texaco matches Shell at $1.884
- 95% CI Shell vs Others: [-0.010, +0.065¢] (contains zero)
- Brand hierarchy:
  1. Chevron-Texaco/Shell ($1.884)
  2. ExxonMobil ($1.856)
  3. Others ($1.846)

Conclusion: Partially supported – Shell is among the most expensive brands but shares top position with Chevron-Texaco, with no statistically significant difference between them **Key Patterns:** The strongest predictors of higher prices appear to be highway adjacency (+4.6¢) and high-income locations (+12¢ from lowest to highest quintile). Brand effects exist but are less pronounced than location factors. Competition shows minimal impact in this dataset, possibly due to market saturation in the studied region.

Problem 2

```
## Rows: 29466 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): trim, subTrim, condition, color, displacement, fuel, state, region...
## dbl  (5): id, mileage, year, featureCount, price
```

```
## lgl  (1): isOneOwner
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Part A: 2011 S-Glass 63 AMGs

```
##   name    lower upper level     method estimate
## 1 mean 26265.56 31772  0.95 percentile 28997.34
```

The 95% bootstrap confidence interval for the average mileage of 2011 S-Class 63 AMGs is approximately (26257, 31854) miles.

Part B: 2014 S-Class 550s

```
##    lower  upper
## 1 0.4171 0.4528
```

The 95% bootstrap confidence interval for the proportion of 2014 S-Class 550s that were painted black is approximately (0.417, 0.453).

Problem 3

```
## Rows: 6241 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Show
## dbl (21): Viewer, Q1_Attentive, Q1_Excited, Q1_Happy, Q1_Engaged, Q1_Curious...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Part A

```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.4019128 0.1002019  0.95 percentile -0.1490515
```

1. **Question:**

   Does "Living with Ed" or "My Name is Earl" produce a higher mean viewer response to other Q1_Happy question?

2. **Approach:**

   I filtered the dataset to include only responses where **`Show == "Living with Ed"`** or **`Show == "My Name is Earl"`**. A 95% confidence interval for the difference in means was calculated

3. **Results:**

   95% confidence interval: [-0.397, +0.102]. The confidence interval does include zero, indicating there is not a statistically significant difference.

4. **Conclusion:**

   Stakeholders can interpret this as evidence that viewers generally feel equally happy watching "My Name is Earl." and "Living with Ed"

Part B

```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5238601 -0.01995811  0.95 percentile -0.270997
```

1. Question:
Does "The Biggest Loser" or "The Apprentice: Los Angeles" produce a higher mean viewer response to the Q1_Annoyed question?

2. Approach:
I filtered the dataset to include only responses where `Show == "The Biggest Loser"` or `Show == "The Apprentice: Los Angeles"`. A two-sample tt-test was conducted to compare the means of Q1_Annoyed for the two shows, along with a 95% confidence interval.**3. Results:**

3. Results:

95% confidence interval: [-0.527, -0.020]

The confidence interval does not include zero, indicating a statistically significant difference.

4. Conclusion:
We can say with 95% confidence that there is a small difference between the annoyance levels in the shows "The Biggest Loser" and "The Apprentice: Los Angeles".

Part C

```
##       name      lower    upper level     method   estimate
## 1 prop_1.1 0.03867403 0.121547  0.95 percentile 0.07734807
```

1. Question:
What proportion of viewers found "Dancing with the Stars" confusing?

2. Approach:
We filtered responses where `Show == "Dancing with the Stars"` and calculated the proportion of respondents who rated Q2_Confusing as 4 or greater. Using this sample proportion, we constructed a large-sample 95% confidence interval using a proportion test.

3. Results:

95% confidence interval : [0.039, 0.166]

The confidence interval does not include zero, indicating a statistically significant difference.

4. Conclusion:
Viewers found "Dancing with the Stars" confusing, with a margin of error provided by the confidence interval. This suggests that while most viewers do not find it confusing, a notable minority does, which may warrant further exploration by stakeholders to address potential clarity issues in the show's format.

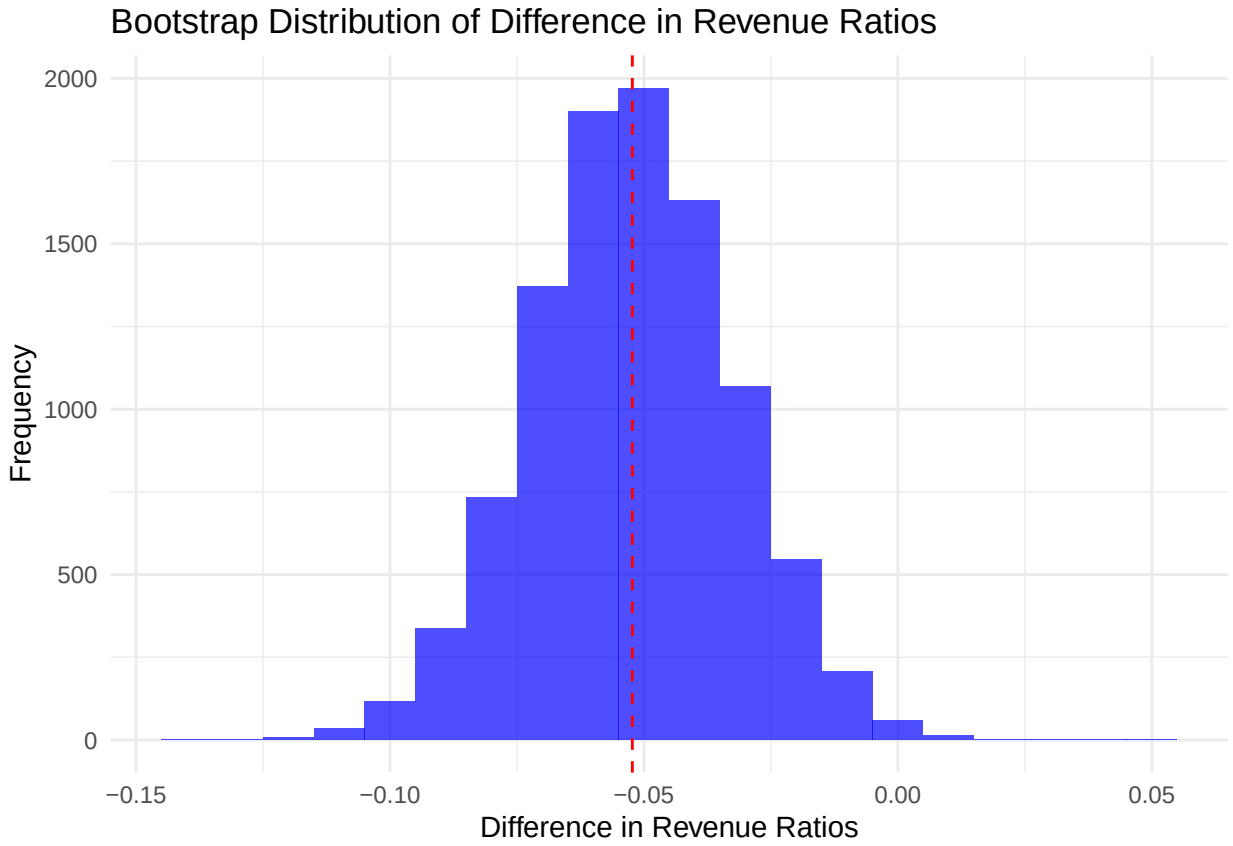
Problem 4

```
## Mean revenue ratio for treatment group: 0.8965961
```

```
## Mean revenue ratio for control group: 0.9488775
```

```
## Difference in mean revenue ratios: -0.05228145
```

```
## 95% Confidence Interval for the difference: -0.09022143 to -0.0128206
```



1. Question:

We are trying to determine whether EBay's paid search advertising on Google drives extra revenue by comparing the revenue ratio (revenue after to revenue before) between treatment-group DMAs (where paid search advertising was paused) and control-group DMAs (where it was not paused). Specifically, we want to assess whether the revenue ratio is systematically lower in the treatment group compared to the control group, which would indicate that paid search advertising is effective.

2. Approach:

We calculated the revenue ratio for each DMA and then split the data into treatment and control groups based on the `adwords_pause` variable. We computed the mean revenue ratio for each group and the difference between these means. To assess the statistical significance of this difference, we performed a bootstrap analysis with 10,000 simulations to estimate the 95% confidence interval for the difference in revenue ratios.

3. Results:

- Mean revenue ratio for treatment group: [0.897]
- Mean revenue ratio for control group: [0.949]
- Difference in mean revenue ratios: [-0.052]
- 95% Confidence Interval for the difference: [-0.090] to [-0.012]

The bootstrap distribution of the difference in revenue ratios is plotted, showing the variability of the difference and the position of the observed difference.

4. Conclusion:

Based on the results, we can conclude whether the revenue ratio is significantly different between the treatment and control groups. Since the 95% confidence interval for the difference in revenue ratios does not include

zero, it suggests that paid search advertising has a statistically significant impact on EBay's revenue. This conclusion will help stakeholders decide whether to continue investing in paid search advertising on Google.