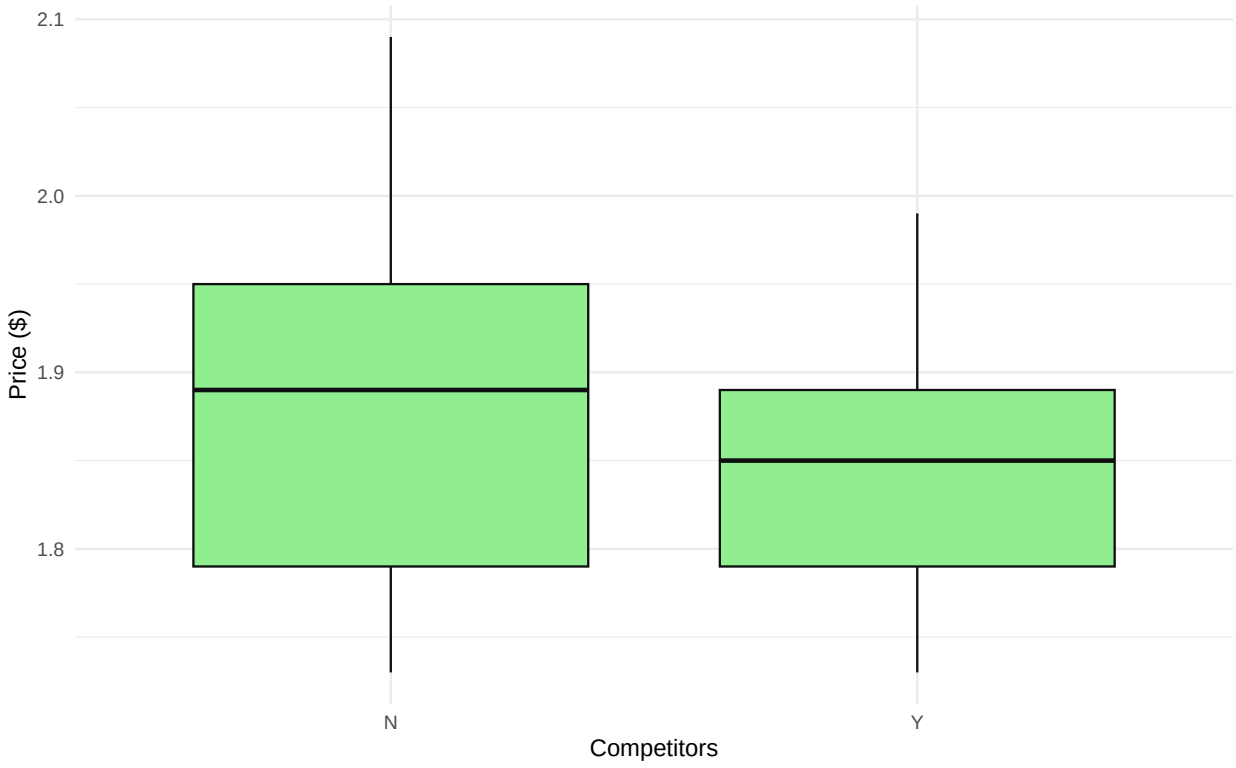
Homework 3

Pranav Belligundu(psb898) - SDS 315 UT Austin

https://github.com/pranav-B21/SDS-315/tree/main/HW%203

Problem 1

Part A - Gas stations charge more if they lack direct competition in sight




```
##       name       lower       upper level     method    estimate
## 1 diffmean -0.05511385 0.007906324  0.95 percentile -0.02348235
```

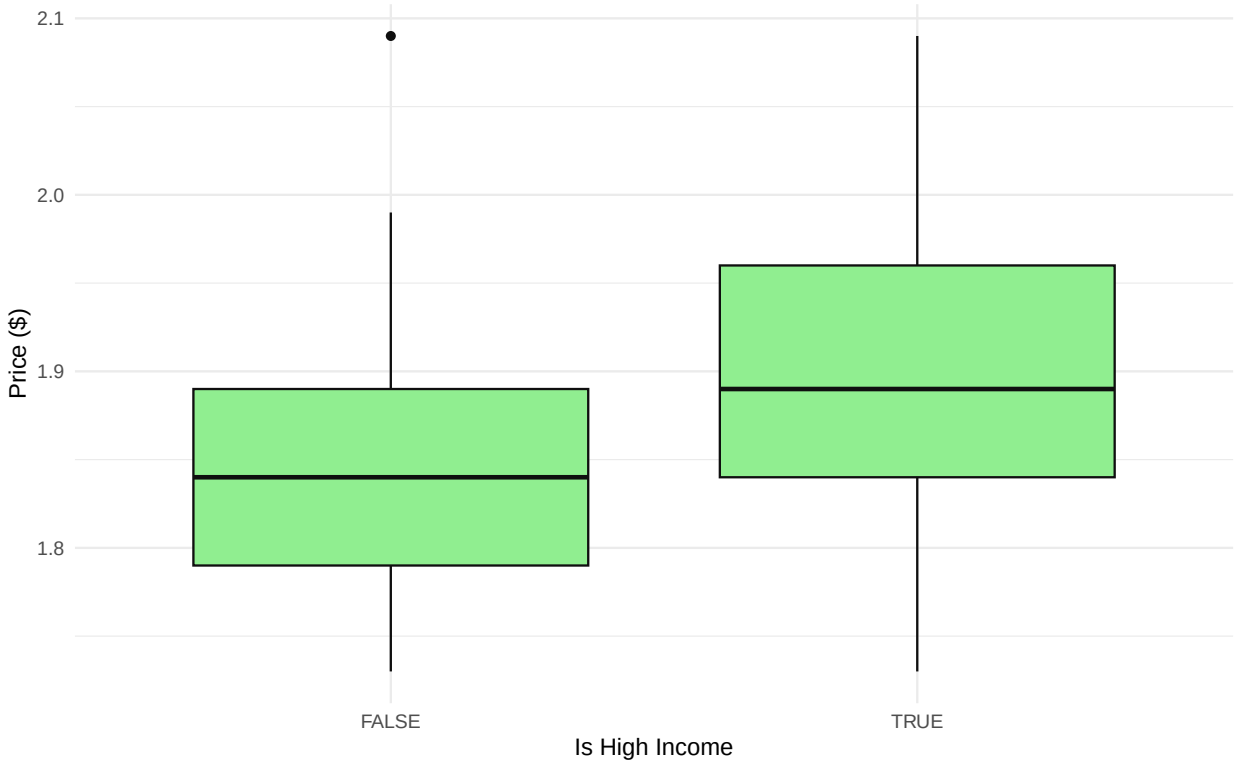
Claim The hypothesis is that gas stations with no direct competitors in sight tend to charge higher prices than those with visible competition. This assumes that stations with no nearby competitors face less price pressure and can increase prices without losing customers.

Evidence The 95% confidence interval for the difference in mean price between stations with and without competitors is (-0.0557, 0.0076), with an estimated mean difference of -0.0235. Additionally, the boxplots show that while gas stations without competitors have a slightly higher median price, there is a lot of overlap in the price distributions. Shows that competitors dont affect much.

Conclusion Since the confidence interval includes zero, there is no statistically significant difference, meaning gas stations without competition do not consistently charge higher prices. This, the data does not support the theory that gas stations charge significantly higher prices when they lack direct competition.

Part B - The richer the area, the higher the gas prices

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12786   37690   52306   56727   70095  128556
```



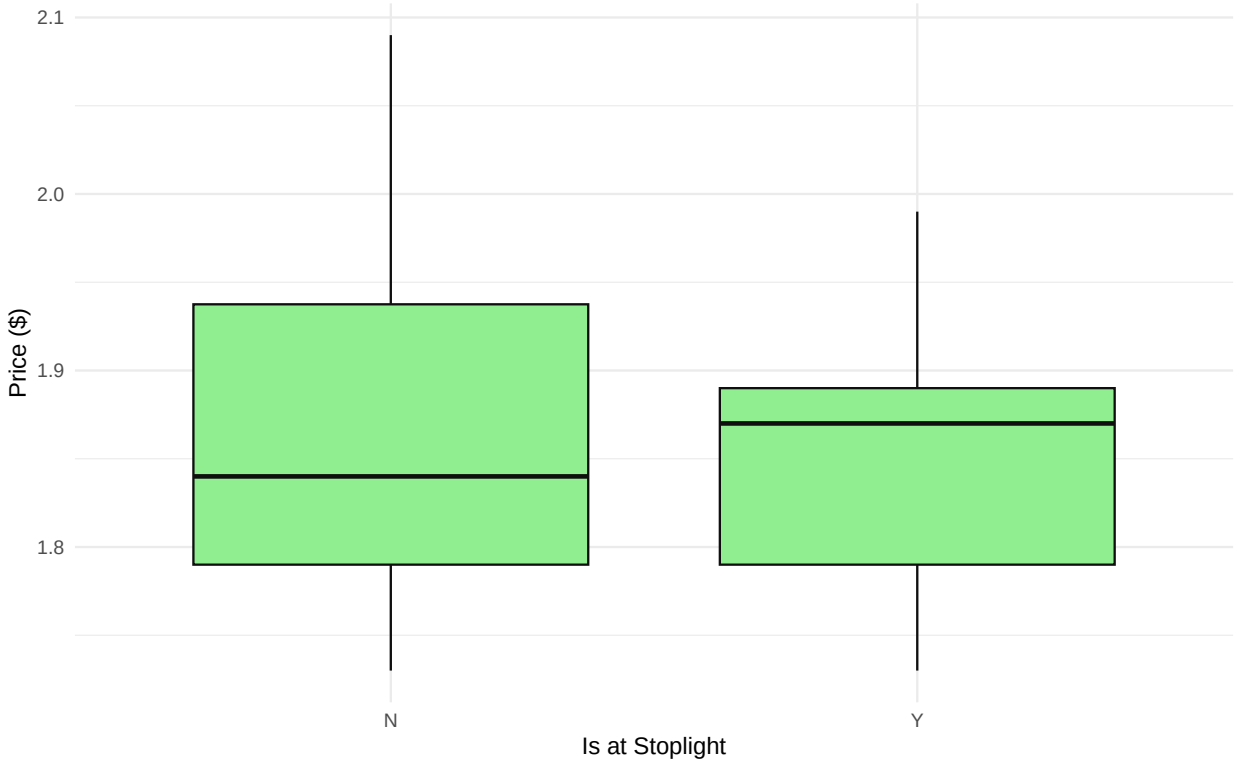

```
##       name      lower      upper level     method   estimate
## 1 diffmean 0.01360513 0.08485937  0.95 percentile 0.04917146
```

Claim The hypothesis is that gas stations located in high-income areas tend to charge higher prices compared to those in lower-income areas, possibly due to greater willingness to pay or higher operating costs.

Evidence The 95% confidence interval for the difference in mean price between stations in high-income and low-income areas is (0.0114, 0.0857), with an estimated mean difference of 0.0492. Since the confidence interval does not include zero, this suggests a statistically significant price difference, meaning stations in high-income areas tend to charge higher prices. Additionally, the boxplots show that gas prices in high-income areas have a higher median and overall distribution compared to lower-income areas.

Conclusion The data supports the theory that gas stations in high-income areas charge higher prices. The statistically significant difference and the clear trend in the boxplots suggest that gas stations in wealthier areas have systematically higher gas prices.

Part C - Gas stations at stoplights charge more.



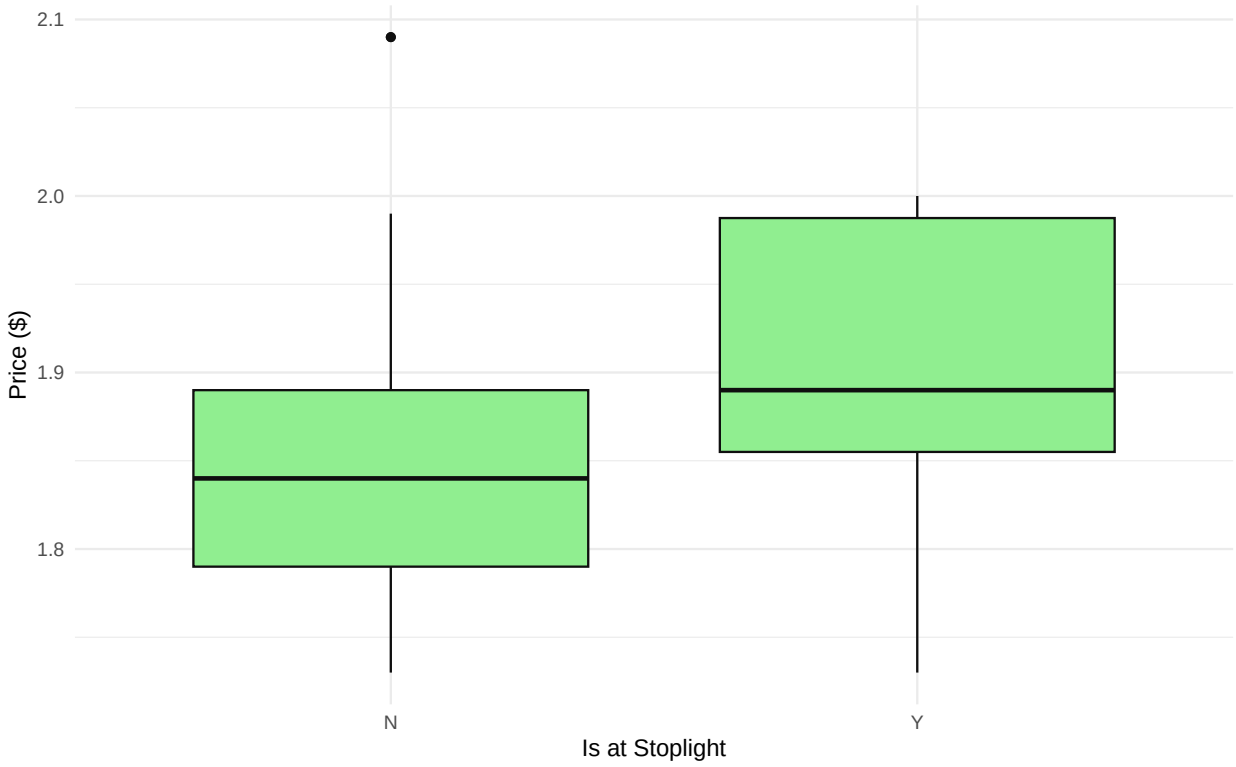

```
##       name       lower      upper level     method     estimate
## 1 diffmean -0.03880661 0.03105879  0.95 percentile -0.003299916
```

Claim The hypothesis is that gas stations located at stoplights charge higher prices possibly because they they benefit due to convenience for drivers who might be more likely to stop and refuel.

Evidence The 95% confidence interval for the difference in mean price between stations at stoplights and those not at stoplights is (-0.0377, 0.0306), with an estimated mean difference of -0.0033. Since the confidence interval includes zero, there is no statistically significant difference, suggesting that gas stations at stoplights do not consistently charge higher prices. Additionally, the boxplots show overlapping distributions, reinforcing the lack of a clear price difference between stations at stoplights and those that are not.

Conclusion The data does not support the theory that gas stations at stoplights charge higher prices. The results indicate that being located at a stoplight does not have a significant impact on gas prices.

Part D - Gas stations with direct highway access charge more



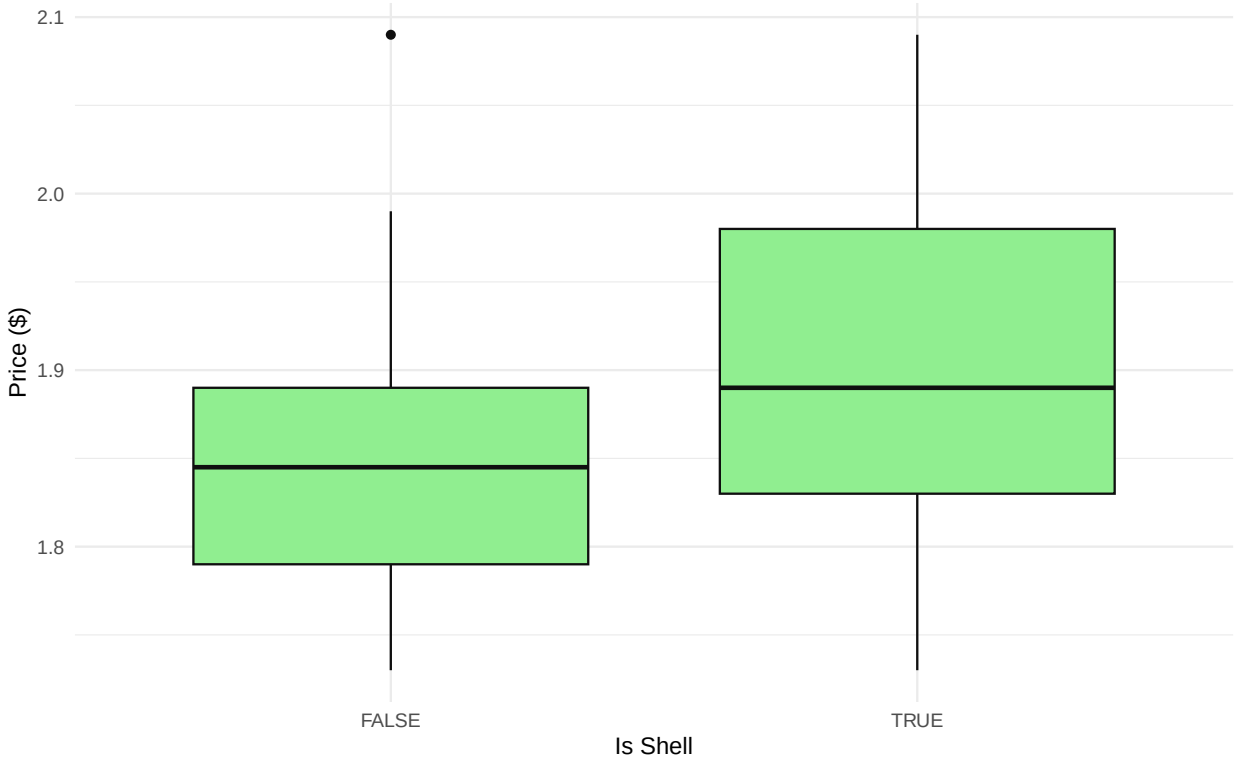

```
##       name       lower      upper level     method  estimate
## 1 diffmean 0.008993721 0.08207301  0.95 percentile 0.0456962
```

Claim The hypothesis is that gas stations located at stoplights charge higher prices due to it being easy to exit and refuel.

Evidence The 95% confidence interval for the difference in mean price between stations at stoplights and those not at stoplights is (0.0090, 0.0811), with an estimated mean difference of 0.0457. Since the confidence interval does not include zero, this suggests a statistically significant price difference, meaning stations at stoplights tend to charge higher prices. Additionally, the boxplots show that gas stations at stoplights have a higher median price and an overall higher price distribution compared to those not at stoplights, further supporting the claim.

Conclusion The data supports the theory that gas stations at stoplights charge higher prices. The statistically significant difference and the boxplot trend suggest that gas stations at stoplights do indeed have higher prices on average.

Part E - Gas stations with direct highway access charge more




```
##       name        lower      upper level     method   estimate
## 1 diffmean -0.009620891 0.06482131  0.95 percentile 0.02740421
```

Claim The hypothesis is that Shell gas stations charge higher prices compared to non-Shell stations.

Evidence The 95% confidence interval for the difference in mean price between Shell and non-Shell stations is (-0.0089, 0.0663), with an estimated mean difference of 0.0274. Since the confidence interval includes zero, thus there is no statistically significant difference. Additionally, the boxplots show some overlap in price distributions, though Shell stations tend to have a slightly higher median price.

Conclusion The data does not provide strong evidence that Shell gas stations charge significantly higher prices. While there is a slight increase in median prices at Shell stations, there is a lack of statistical significance that can be due to random variance.

Problem 2

Part A

Filter the data set down to include only those cars where year == 2011 and trim == "63 AMG". Based on these 116 cars, compute a 95% bootstrap confidence interval for the average mileage of 2011 S-Class 63 AMGs that were hitting the used-car market when this data was collected.

```
##   name    lower    upper level     method estimate
## 1 mean 26303.95 31770.74  0.95 percentile 28997.34
```

Part B

Filter the data set down to include only those cars where year == 2014 and trim == "550". Based on this sample of 2889 cars, compute a 95% bootstrap confidence interval for the proportion of all 2014 S-Class 550s that were painted black. Hint: you might find this easiest if you use mutate to first define a new variable, isBlack, that is either TRUE or FALSE depending on whether the car is black.

```
##        name     lower     upper level     method  estimate
## 1 prop_TRUE 0.4167532 0.4531066  0.95 percentile 0.4347525
```

Problem 3

Part A

```
##       name      lower     upper level     method   estimate
## 1 diffmean -0.3990928 0.1025844  0.95 percentile -0.1490515
```

Question: Consider the shows "Living with Ed" and "My Name is Earl." Who makes people happier: Ed or Earl?

Approach: Used bootstrapping to estimate the difference in mean Q1_Happy responses between the two shows. Specifically, we performed 10,000 resamples and calculated the difference in means for each resample. A 95% confidence interval was then constructed for this difference.

Results: The 95% confidence interval for the difference in mean happiness ratings (Living with Ed - My Name is Earl) is [-0.3892, 0.1005], with an estimated mean difference of -0.1491.

Conclusion: Since this confidence interval includes zero, we do not have strong statistical evidence that one show produces a significantly higher mean happiness response than the other.

Part B

```
##       name      lower       upper level     method  estimate
## 1 diffmean -0.5244622 -0.01747709  0.95 percentile -0.270997
```

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

Approach: Used bootstrapping to estimate the difference in mean Q1_Annoyed responses between the two shows. We performed 10,000 resamples and calculated the difference in means for each resample. A 95% confidence interval was then constructed for this difference.

Results: The 95% confidence interval for the difference in mean annoyance ratings (The Biggest Loser - The Apprentice: Los Angeles) is [-0.5241, -0.0127], with an estimated mean difference of -0.2710.

Conclusion: Since the confidence interval does not include zero and is entirely negative, this suggests that The Apprentice: Los Angeles has a significantly higher mean annoyance rating than The Biggest Loser.

Part C

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

Question: What proportion of American TV watchers would be expected to give a response of 4 or greater to the Q2_Confusing question for Dancing with the Stars?

Approach: Calculated the proportion of viewers who rated the show as confusing and used bootstrapping (10,000 resamples) to construct a 95% confidence interval for this proportion.

Results: The estimated proportion of viewers who found Dancing with the Stars confusing is 0.0773 (7.73%). The 95% confidence interval for this proportion is [0.0387, 0.1160].

Conclusion: This means we estimate that between 3.87% and 11.60% of all American TV viewers would agree or strongly agree that Dancing with the Stars is confusing. Based on this analysis, a relatively small but notable proportion of viewers (around 7.73%) found Dancing with the Stars confusing.

Problem 4

```
##       name       lower      upper level     method    estimate
## 1 diffmean -0.09080845 -0.0138132  0.95 percentile -0.05228145
```

Question: Does EBay's paid search advertising on Google generate additional revenue, or would the company see similar revenue levels by relying on organic search results?

Approach: Calculated the revenue ratio (revenue after / revenue before) for each DMA and compared the treatment group to the control group. Using bootstrapping with 10,000 Monte Carlo simulations, we estimated the difference in mean revenue ratios between these two groups and constructed a 95% confidence interval.

Results: The estimated mean difference in revenue ratio (Treatment - Control) is -0.0523. The 95% confidence interval for this difference is [-0.0910, -0.0138].

Conclusion: These results provide statistical evidence that paid search advertising contributes to increased revenue for EBay. Since the confidence interval does not include zero and is fully negative, this indicates that turning off paid ads had a significant negative impact on revenue.