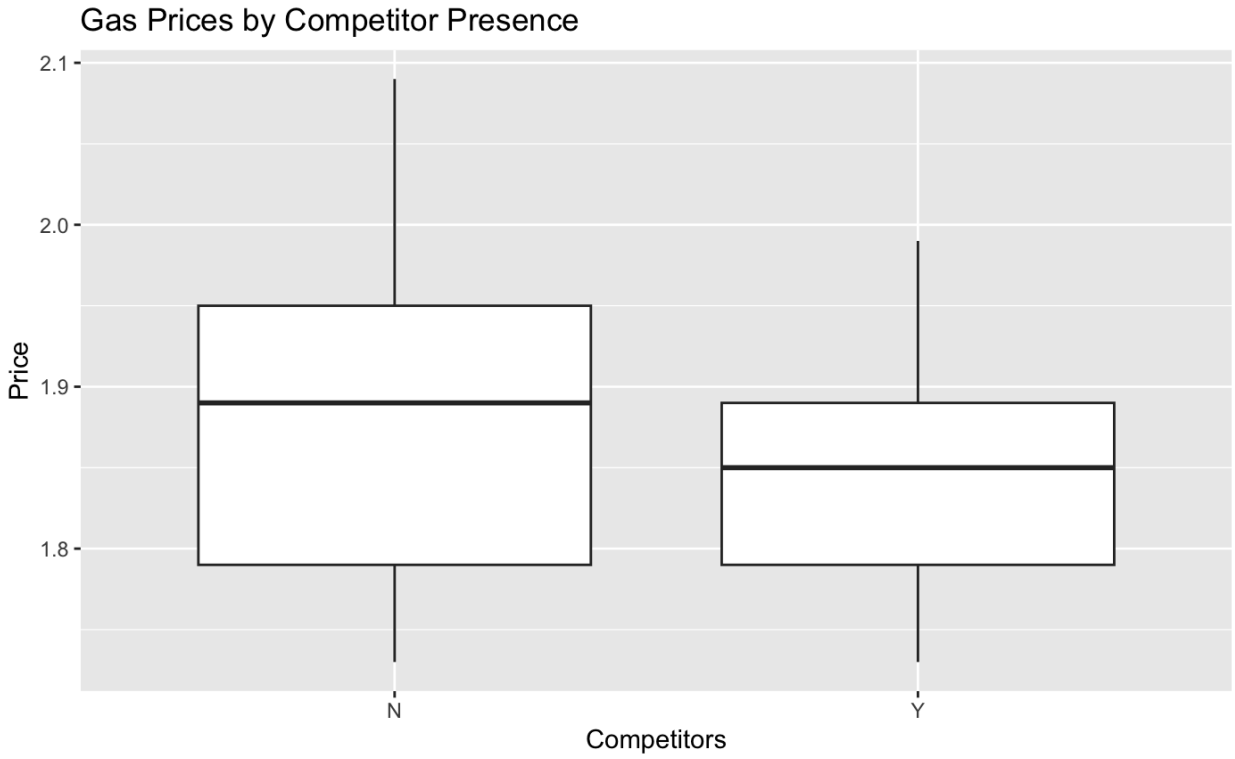
Soumil Asanare
Ssa2958
Github: https://github.com/Soumil-A/HW3SDS.git

Problem 1

Theory A

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

Evidence:




```
                 2.5 %     97.5 %
(Intercept)   1.85335937 1.89840534
CompetitorsY -0.05549353 0.00852882
```

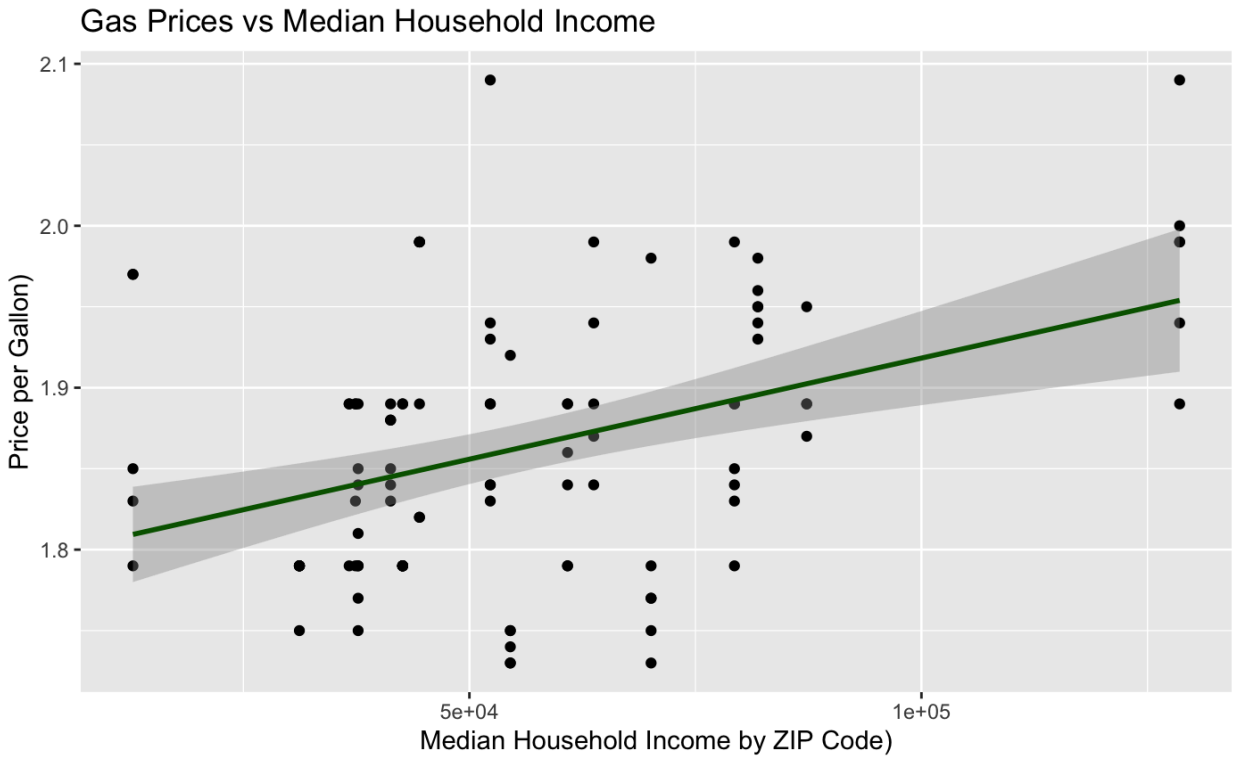
After calculating the confidence intervals there is no difference between the two. The differences in the two groups (price and competitors) is somewhere between -0.055 and 0.008. This means that we are 95% confident the confidence intervals overlap and there is no difference between the two groups.

Conclusion:

With 95% confidence, the price difference between gas stations with and without competitors is between -0.0555 and 0.008 percent. The two groups do not vary statistically because 0 is in the interval. The data does not support the theory.

Theory B

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




```
                2.5 %       97.5 %
(Intercept) 1.757499e+00 1.829386e+00
Income      6.713493e-07 1.825333e-06
```

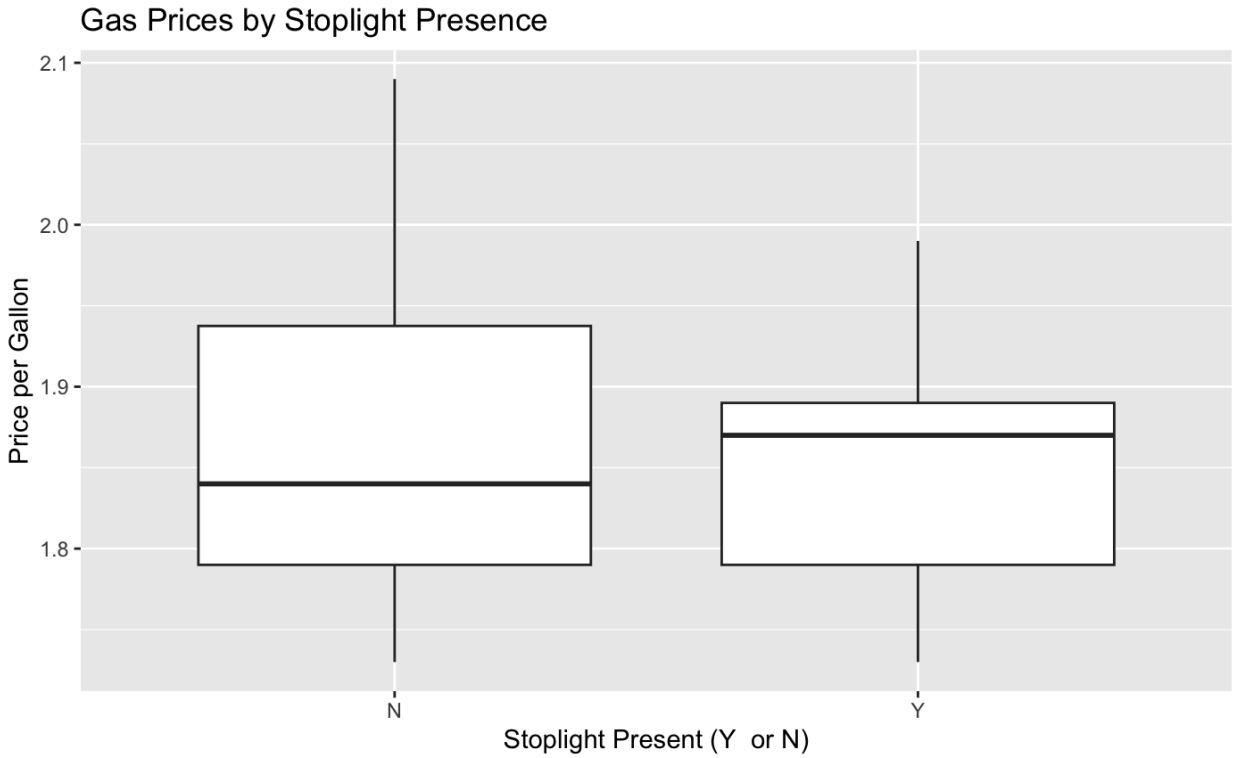
The shown effect of household income on gas price is between 0.0000006713 and 0.000001853 per dollar of income, with 95% confidence.

Conclusion:

The confidence interval is completely positive and doesn't have 0 shows that there is a slight correlation between local income and gas prices. With 95% confidence, the impact of zip code income on gas station prices ranges from 0.000000671 to 0.00000183.

Theory C

Theory: Gas stations at stoplights charge more.




```
                2.5 %     97.5 %

(Intercept)  1.83995048 1.89268110

StoplightY  -0.03668276 0.03008292
```

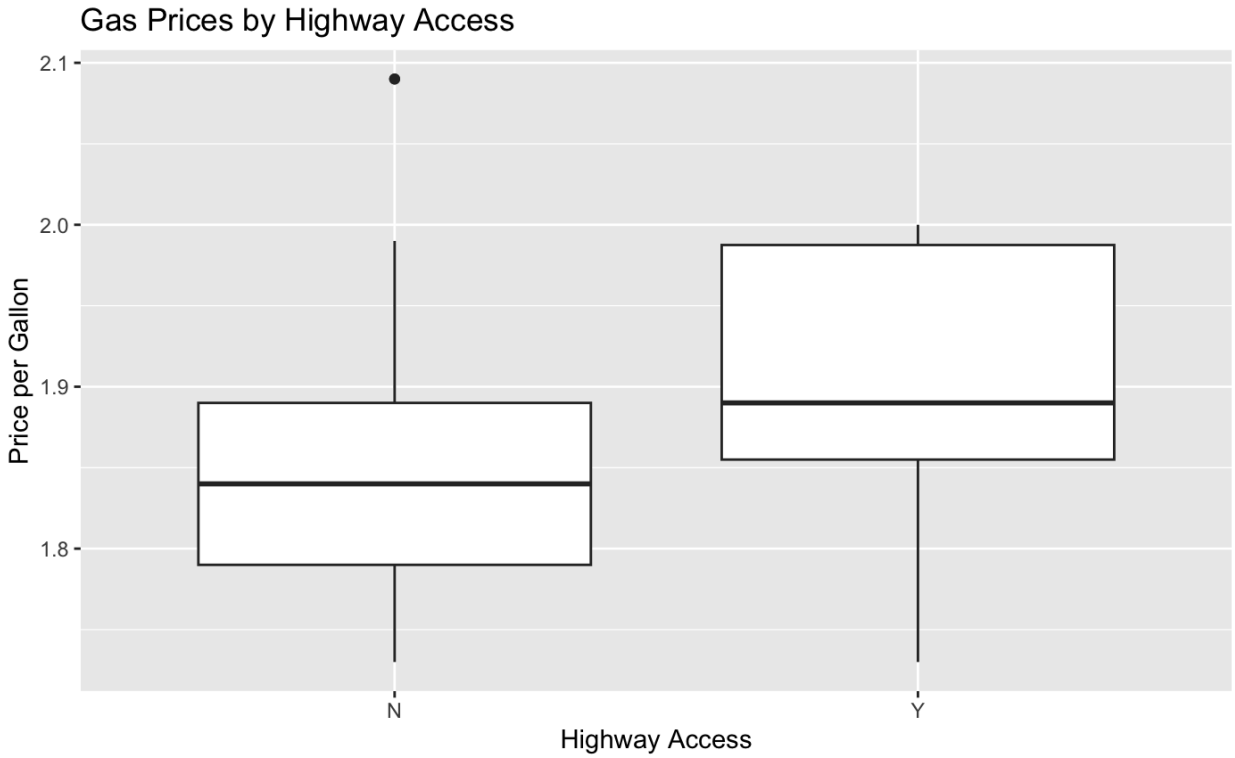
The estimated difference in gas prices between gas stations with a stoplight and those without based on the boxplot and data is between -0.0367 and 0.0301 dollars, with 95% confidence.

Conclusion:

There is no statistically significance between the two groups, because the confidence interval of -0.0367 - 0.0301 includes 0. The theory that gas stations at stoplights charge more is not supported by the data.

Theory D

Theory: Gas stations with direct highway access charge more.




```
                2.5 %     97.5 %
(Intercept) 1.836515952 1.87209164
HighwayY    0.007583242 0.08380916
```

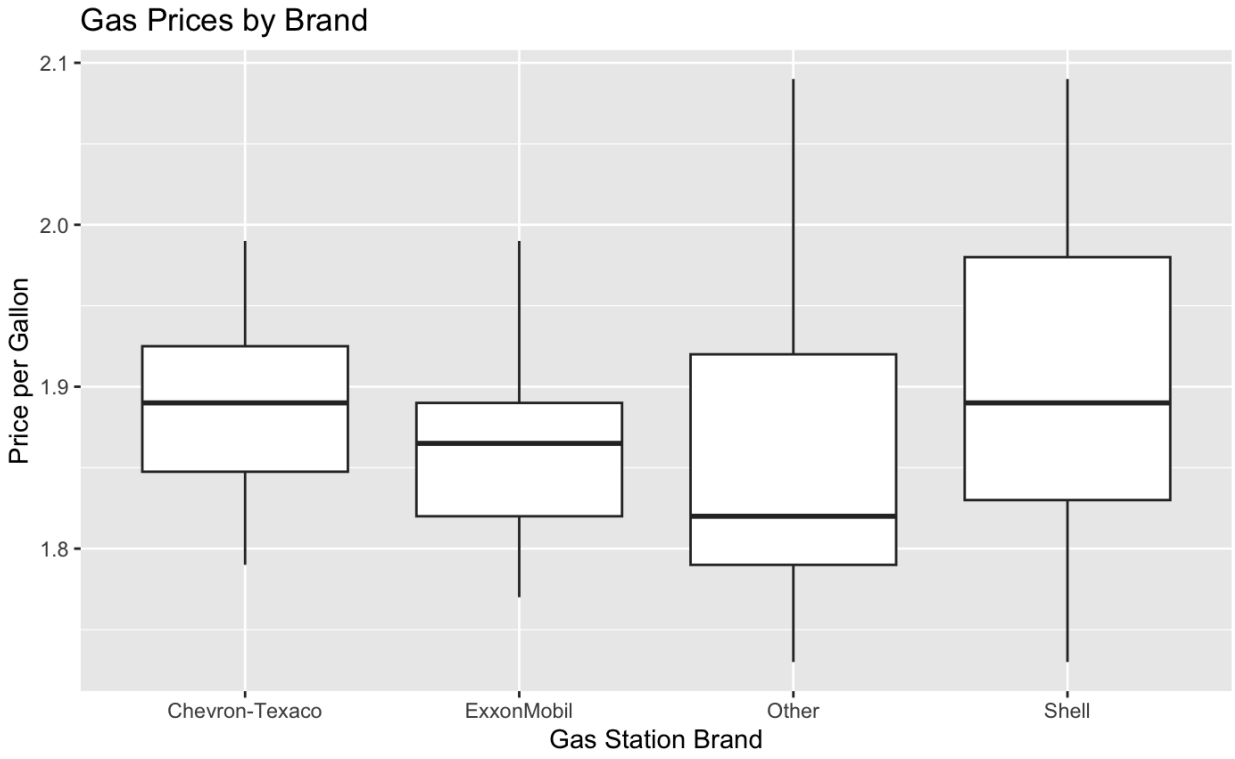
The estimated difference in gas prices between gas stations with highway access and those without is between 0.00758 and 0.08381 dollars, with 95% confidence.

Conclusion:

There is a small statistical difference between the two groups, as shown by the confidence interval (0.00758, 0.08381), which is all positive and doesn't include 0. The theory that gas stations with direct access to the roadway charge higher prices is supported by the data.

Theory E

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




```
                    2.5 %      97.5 %
(Intercept)      1.84433590 1.924414096
BrandExxonMobil -0.08655766 0.030664806
BrandOther      -0.08571223 0.008390801
BrandShell      -0.05045789 0.049294092
```

The confidence interval for Shell's price difference compared to other brands is between -0.0505 and 0.0493 dollars, with 95 % confidence.

Conclusion

There is no statistically significant difference between the two groups, as shown by the confidence interval because it has 0. The theory that shell gas stations charge more than non-shell brands is supported by the data.

Problem 2

Part A

name <chr>	lower <dbl>	upper <dbl>	level <dbl>	method <chr>	estimate <dbl>
mean	26292.13	31846.21	0.95	percentile	28997.34

1 row

I am 95% confident that the average mileage of 2011 S-Class 63 AMGs has the lower bound and upper bound of 26292 and 31846 miles respectively.

Part B

name <chr>	lower <dbl>	upper <dbl>	level <dbl>	method <chr>	estimate <dbl>
prop_TRUE	0.4164071	0.453098	0.95	percentile	0.4354448

1 row

I am 95% confident that the proportion of 2011 S-Class 63 AMGs that are black has the lower bound and upper bound of 0.4164 and 0.4531 miles respectively.

Problem 3

Part A

Question: Is there evidence that one show consistently produces a higher mean Q1_Happy response among viewers?

Approach: Filter the dataset, then take a 95% confidence interval which takes the diffrence in the confidence intervals of the 2 shows.

```
                          2.5 %    97.5 %

(Intercept)           3.7252500 4.1284085

ShowMy Name is Earl -0.3988754 0.1007724
```

The confidence interval for the difference in happy scale between the two shows is -0.3988 and 0.1007 with 95% confidence.

Conclusion:

We are 95% confident that there is no statistical proof that one show makes viewers happier than the other because the confidence interval contains 0. Although My Name is Earl's projected happiness score is lower than Living with Ed's, the difference is not statistically significant.

Part B

Question: Does the show The Biggest Loser or The Apprentice: Los Angeles make people more annoyed?

Approach: Take a 95% confidence interval from that which takes the difference in the confidence intervals of the 2 shows.

```
                             2.5 %      97.5 %

ShowThe Biggest Loser -0.5273332 -0.01466086
```

The confidence interval for the difference in happy scale between the two shows is -0.5273 and -0.0147, with 95% certainty.

Conclusion

We are 95% convinced that there is a statistical difference between the two shows' degrees of annoyance because the confidence interval is entirely negative and does not contain 0. The Apprentice: Los Angeles seems to have a greater mean annoyance rating than The Biggest Loser, based on the negative confidence interval.

Part C

Question: Based on a sample of respondents who watched Dancing with the Stars, what proportion of American TV watchers would we expect to report being confused by the show?

Approach: Take a 95% confidence interval from that which takes the difference in the confidence intervals of the 2 shows and how confused by the shows.

```
                2.5 %    97.5 %

(Intercept) 0.03805777 0.1166384
```

The confidence interval for the difference in happy scale between the two shows is 0.0380 and 0.1166, with 95% certainty.

Conclusion:

According to our data, there is a 95% chance that American TV viewers, ranging from 0.038 to 0.117, will answer "4" or higher to the Q2_Confusing Question. We are 95% convinced that the confusing Dancing with the Stars levels have a statistical significance because the confidence interval is positive and does not contain 0.

Problem 4

Question: Does the extra traffic brought to our site from paid search results justify the cost of the ads themselves?

Approach: Compute the revenue ratio for each DMA. Fit a linear model to estimate the difference in revenue ratio between treatment and control groups. Make a confidence interval for the difference.

```
                    2.5 %      97.5 %

(Intercept)     0.92743669  0.97031837

adwords_pause -0.09380147 -0.01076144
```

The confidence interval for the difference in adwords pause is -0.0938 and -0.0107 with 95% certainty.

Conclusion:

According to the data, the advertisements paused group has a lower revenue ratio than the other group. With 95% confidence, the difference in the two groups is between -0.0938 and -0.0107. We are 95% certain that the revenue ratio between the treatment group and control group is statistically significant because 0 is not in the confidence interval and it is negative.