Homework3SDS315

Kapil Taspa

2025-02-11

UT EID: kt27955

GitHub Repo: https://github.com/ktaspa/Homework3SDS315

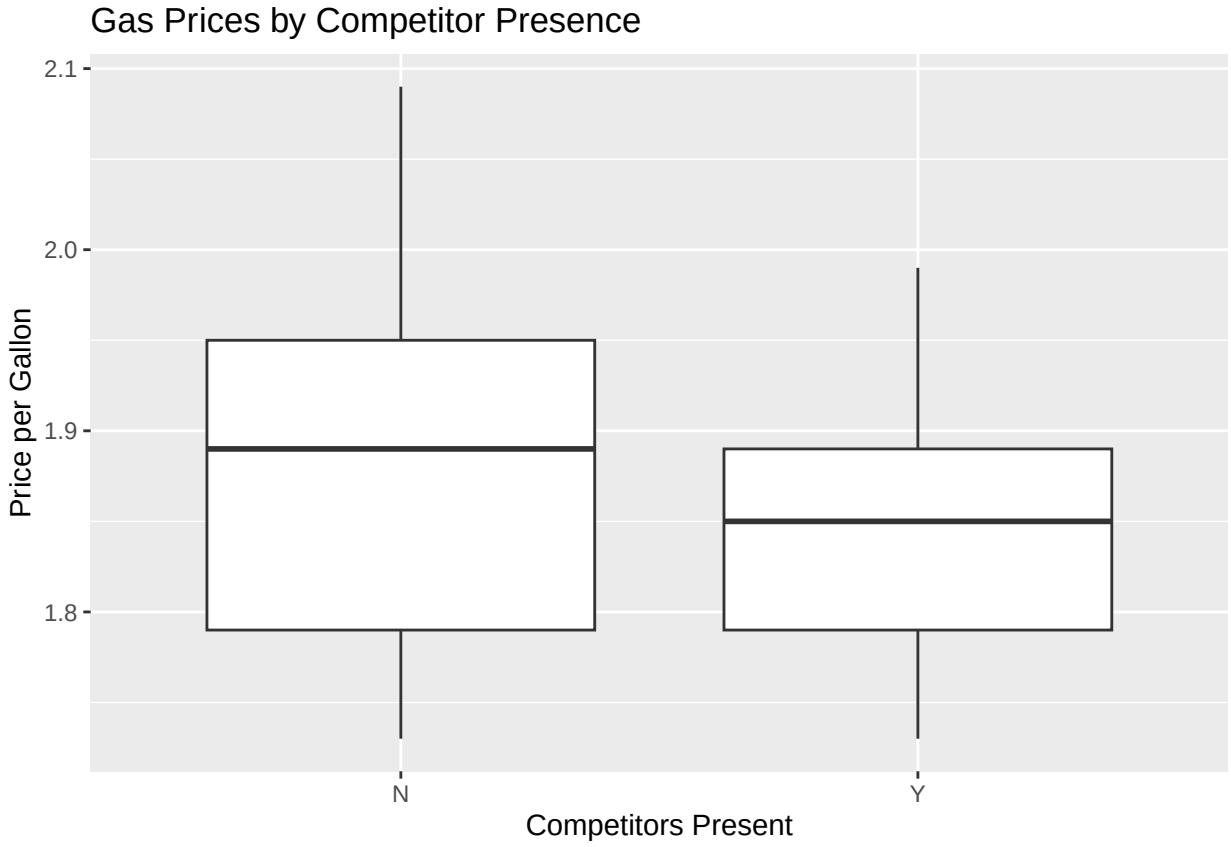
Problem 1

gasprices.csv

Theory A

Gas Stations charge more if they lack direct competition in sight.

Evidence




```
##                    2.5 %     97.5 %
## CompetitorsY -0.05549353 0.00852882
```

The data seems to support the theory, but after calculating the confidence intervals we know there is no difference between the two. The differences in the two groups is somewhere between -0.055 and 0.008. This means that the confidence intervals overlap and there is no significant difference between the two groups with a 95% certainty.

Conclusion

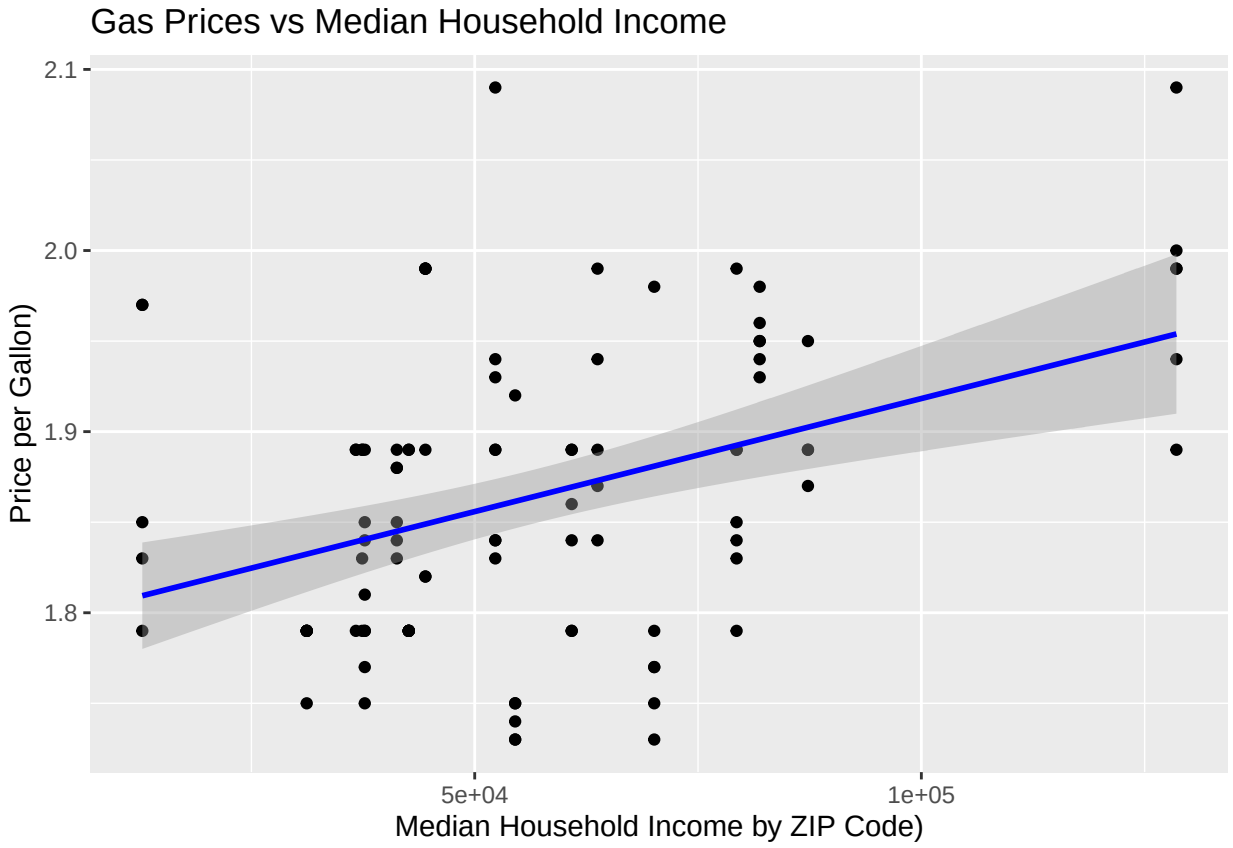
The difference in price between gas stations with competitors present or not present is somewhere between -0.0555 and 0.008, with 95% confidence. Because 0 is in the interval, there is no statistical difference between the two groups. The theory is unsupported by the data.

Theory B

The richer the area, the higher the gas prices.

Evidence

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
##                2.5 %       97.5 %
## Income 6.713493e-07 1.825333e-06
```

A linear regression model was fitted to predict gas prices based on median household income. The estimated effect of income on gas price is extremely small—between 0.000000671 and 0.00000183 per additional dollar of income, with 95% confidence.

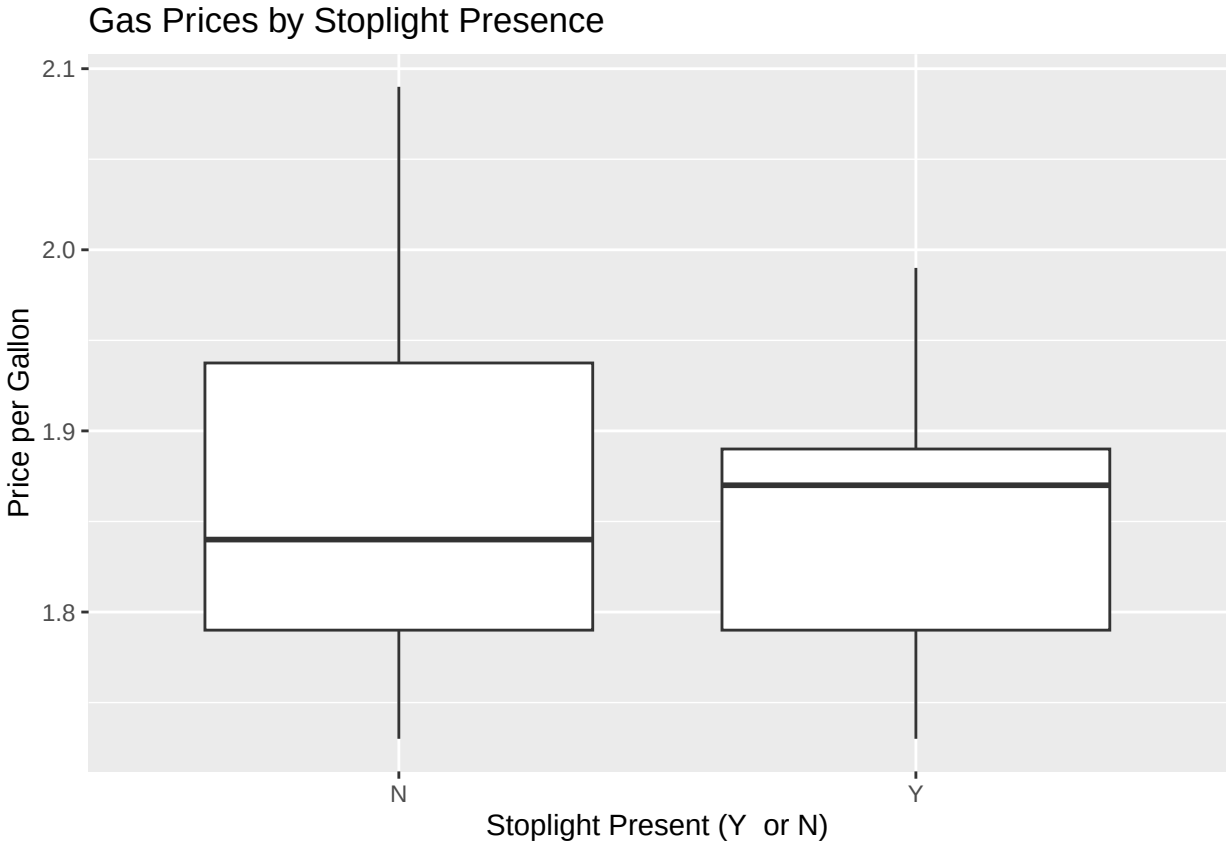
Conclusion

The confidence interval does not include 0 and is all positive so this means there is a relationship between the income of the area and gas prices although it is small. The effect of income of zip code on the price of gas stations is somewhere between 0.000000671 and 0.00000183, with 95% certainty.

Theory C

Gas stations at stoplights charge more.

Evidence




```
##                   2.5 %     97.5 %
## StoplightY -0.03668276 0.03008292
```

A linear regression model was fitted to predict gas prices based on the presence of a stoplight. The estimated difference in gas prices between gas stations with a stoplight and those without is between -0.0367 and 0.0301 dollars, with 95% confidence.

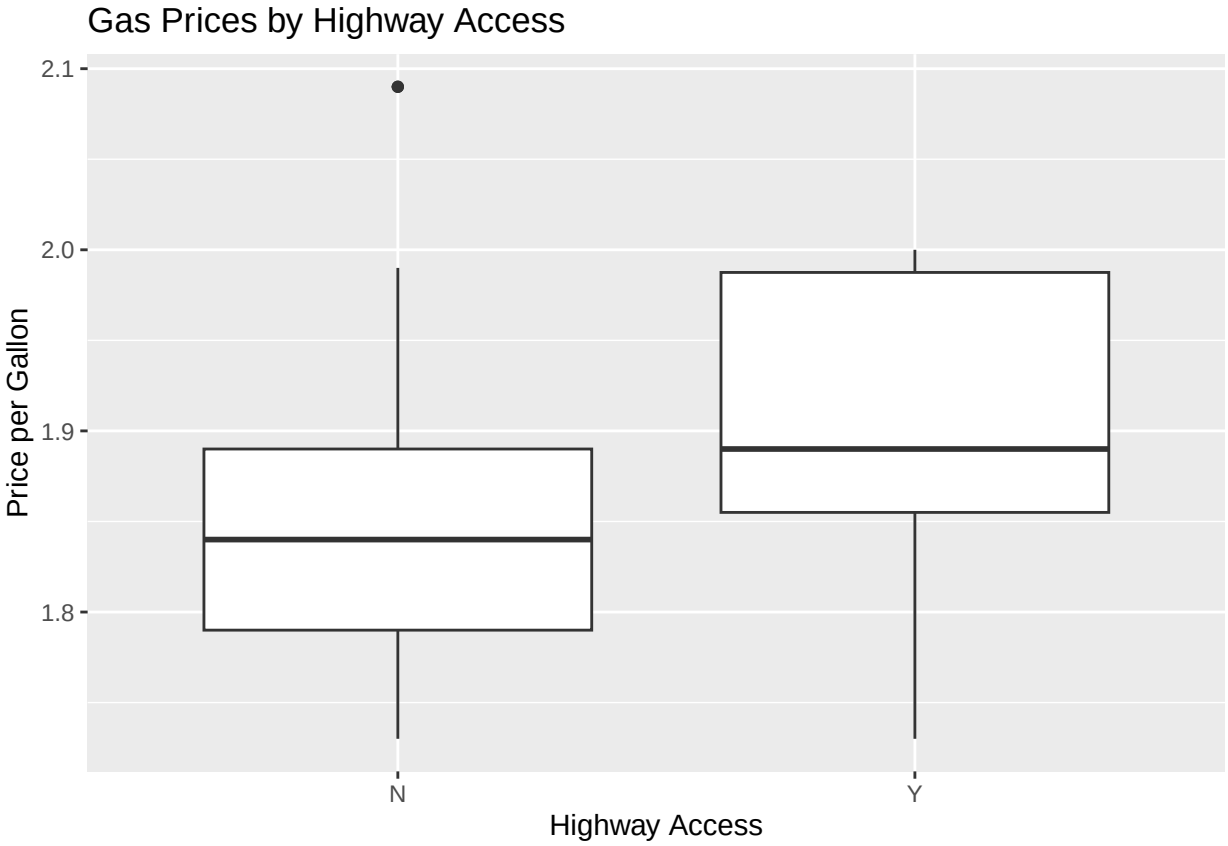
Conclusion

The confidence interval includes 0 which tells us that there is no statistical difference between the two groups. The data does not support the theory that gas stations at stoplights charge more.

Theory D

Gas stations with direct highway access charge more.

Evidence




```
##                2.5 %     97.5 %
## HighwayY 0.007583242 0.08380916
```

A linear regression model was fitted to predict gas prices based on highway access. The estimated difference in gas prices between gas stations with highway access and those without is between 0.0076 and 0.0838 dollars, with 95% confidence.

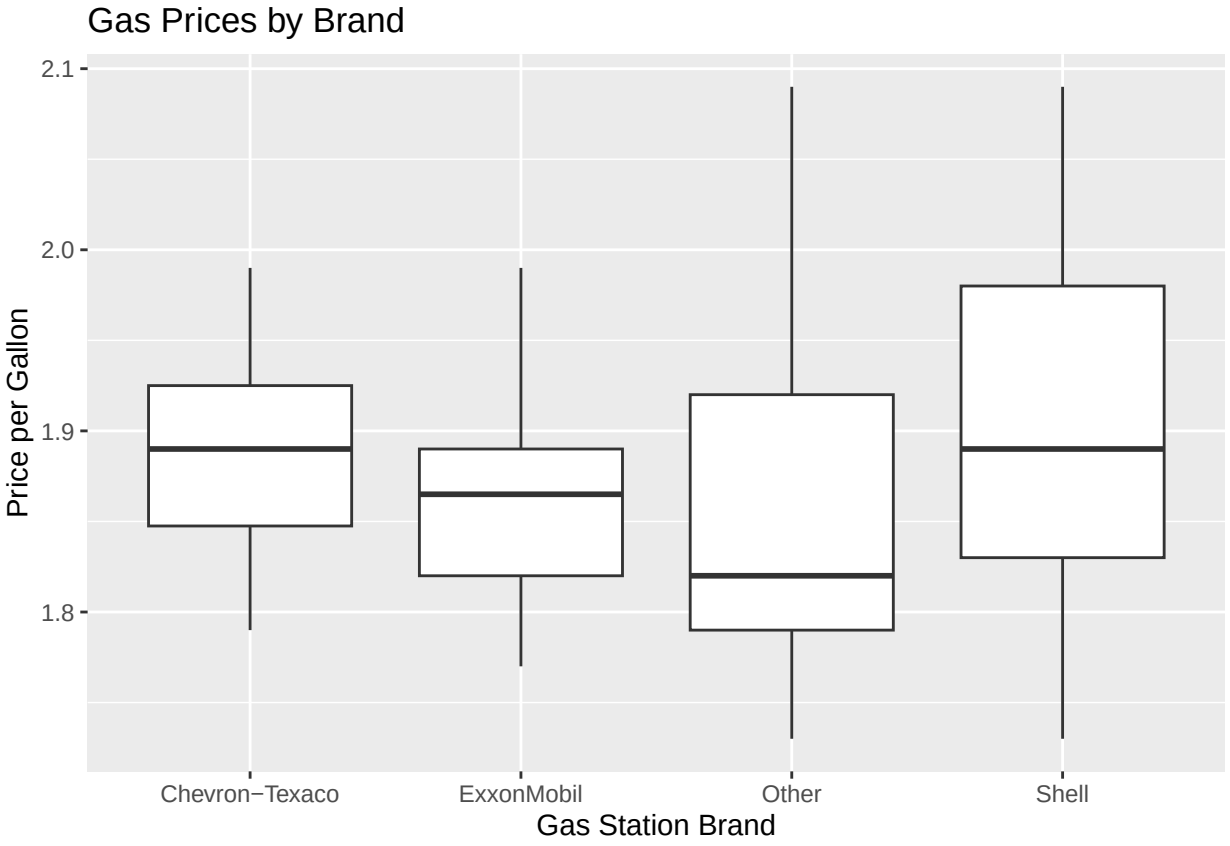
Conclusion

The confidence interval does not include 0 and is all positive which tells us that there is a statistical difference between the two groups although it may be a small difference. The data does support the theory that gas stations that have direct access to the highway charge more.

Theory E

Shell charges more than all other non-Shell brands.

Evidence




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

A linear regression model was fitted to predict gas prices based on gas station brand. The confidence interval for Shell's price difference compared to other brands is between -0.0505 and 0.0493 dollars, with 95 % confidence.

Conclusion

The confidence interval includes 0 which tells us that there is not a statistical difference between the two groups. The data does support the theory that shell gas stations charge more than non-shell brands of gas stations.

Problem 2

sclass.csv

Part A

```
##   name    lower    upper level     method estimate
## 1 mean 26276.03 31805.03  0.95 percentile 29317.04
```

Part B

```
##        name     lower    upper level     method estimate
## 1 prop_TRUE 0.4167532 0.453098  0.95 percentile 0.423676
```

Problem 3

Part A

Question

Does the show Living with Ed or My Name is Earl make people happier?

Approach

filter the dataset to only include living with ed and the my name is earl fitted a linear model then took a 95% confidence interval from that which takes the difference in the confidence intervals of the 2 shows.

Results

```
##                            2.5 %    97.5 %
## ShowMy Name is Earl -0.3988754 0.1007724
```

The confidence interval for the difference in happy scale between the two shows is -0.3989 and 0.1008, with 95% certainty.

Conclusion

The confidence interval contains 0 which means that we are 95% certain that there is no statistical evidence that one show makes viewers more happy than the other. While My Name is Earl has a slightly lower estimated happiness score compared to Living with Ed the difference is small and not statistically significant.

Part B

Question

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

Approach

filter the dataset to only include the biggest loser and the apprentice. fitted a linear model then took a 95% confidence interval from that which takes the difference in the confidence intervals of the 2 shows.

Results

```
##                              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

The confidence interval does not contains 0 and is all negative which means that we are 95% certain that there is a statistical difference between the annoyance levels of the two shows. The negative confidence interval suggests that The Apprentice: Los Angeles has a higher mean annoyance rating compared to The Biggest Loser.

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

fitted a linear model then took a 95% confidence interval from that which takes the difference in the confidence intervals of the 2 shows.

Results

```
##                   2.5 %    97.5 %
## (Intercept) 0.03805777 0.1166384
```

```
## (Intercept)
##  0.07734807
```

The confidence interval for the difference in happy scale between the two shows is 0.03806 and 0.11664, with 95% certainty.

Conclusion

Based on our sample, a proportion of somewhere between 0.038 and 0.117 of American TV watchers would be expected to give a response of 4 or higher to the Q2_Confusing Question, with 95% certainty. The confidence interval does not contain 0 and is all positive which means that we are 95% certain that there is a statistical significance of the confusing levels of Dancing with the Stars.

Problem 4

Question

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

Approach

created a new variable rev_ratio. fitted a linear model then took a 95% confidence interval from that which takes the difference in the confidence intervals of the 2 groups.

Results

```
##                      2.5 %      97.5 %
## (Intercept)    0.92743669  0.97031837
## adwords_pause -0.09380147 -0.01076144
```

The confidence interval for the difference in adwords pause between the two groups is -0.0938 and -0.0108, with 95% certainty.

Conclusion

Based on our sample, the revenue ratio in the treatment group, where ads are paused, is lower than in the control group. The difference is somewhere between -0.0938 and -0.0108, with 95% certainty. 0 is not in the confidence interval and it's all negative so we are 95% certain that there is a statistical significance in the revenue ratio between the treatment group and control group.