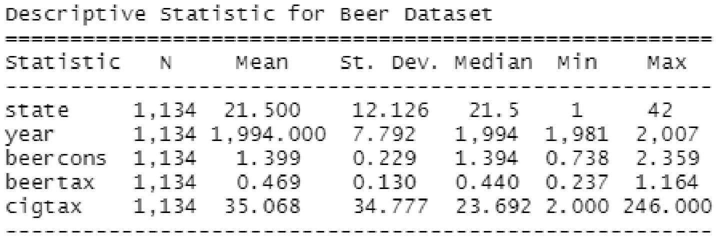
Name	Student ID Number
Do Nhat Anh Ha	1194034

Question 1:

As from the figure below, a typical state, on average, has around **1.399 gallons of beer sold per capita**. Moreover, the average beer tax in a typical state **is $0.469 per gallon** and the average cigarette tax in a typical observation is **$35.068 per pack**. For beer sales, the min (**1,981 gallons per capita**) and the max (**2,007 gallons per capita**) aren't too far away, and the mean (**1.399 gallons per capita**) and median (1.394 gallons per capita) are relatively close which possibly indicates a symmetrical distribution. In contrast, beer tax has a mean ($0.469 per gallon) that is greater than the median ($0.440 per gallon), which might possibly indicate right skew. Also, the range ($0.927 per gallon) between min ($0.237 per gallon) and max value ($1.164 per gallon) is quite considerable, hence there is more extreme value on the right of the distribution of beer tax. For cigarette tax, the mean ($35.068 per pack) is also greater than the median ($34.777 per pack). It's possible that the distribution of cigarette tax is right skew. Also, the min ($2 per pack) is closer to the central tendency than the max ($246 per pack) which shows that there are more extreme values toward the right tail of the distribution.



```
Descriptive Statistic for Beer Dataset
=======================================================
Statistic   N      Mean     St. Dev. Median  Min    Max
-------------------------------------------------------
state     1,134  21.500     12.126   21.5    1      42
year      1,134 1,994.000  7.792    1,994  1,981  2,007
beercons  1,134   1.399     0.229    1.394  0.738  2.359
beertax   1,134   0.469     0.130    0.440  0.237  1.164
cigtax    1,134  35.068     34.777  23.692 2.000 246.000
-------------------------------------------------------
```

Question 2:

For beer sales (gallons per capita), the 95% CI is:

$$[1.386, 1.412]$$

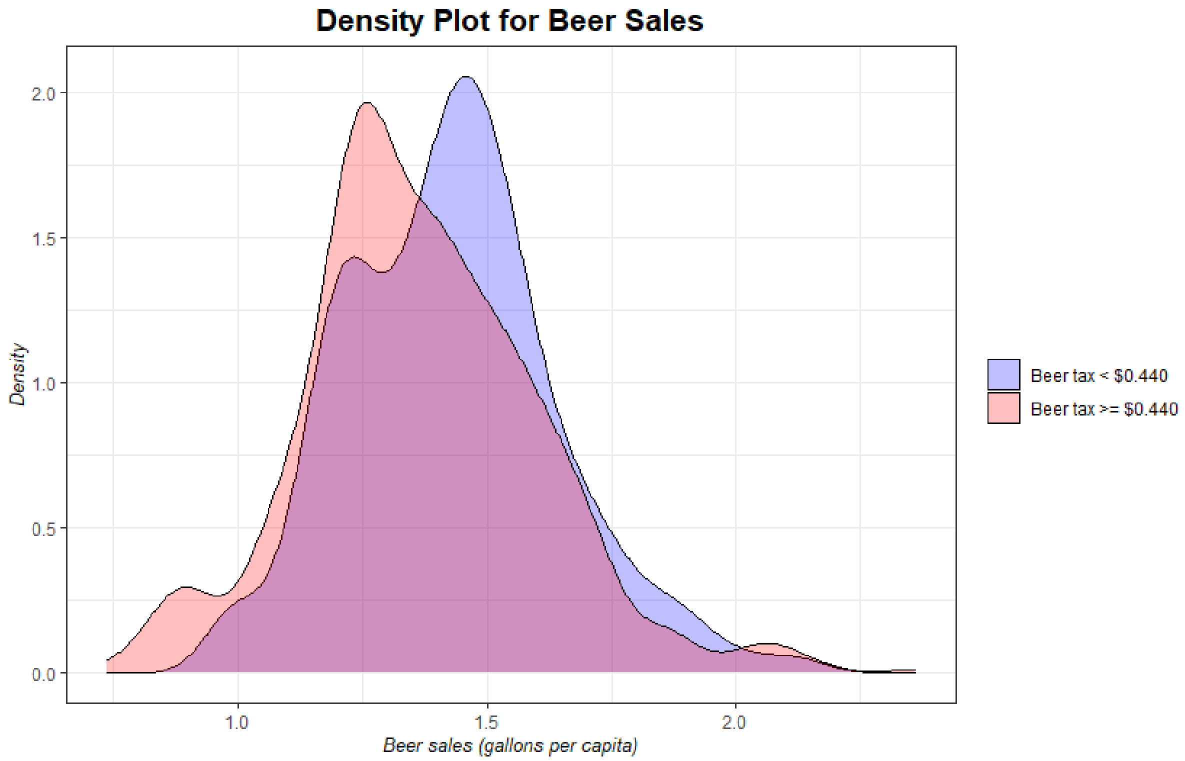
For beer tax (dollars per gallon), the 95% CI is:

$$[0.461, 0.476]$$

For cigarette tax (dollars per pack), the 95% CI is

$$[33.044, 37.092]$$

Question 3:



From the graph above, we can see that the mean for beer sales when beer tax < $0.440 per gallon is greater than the mean for beer sales when beer tax >= $0.440 per gallon. We can also see that the probability mass for beer sales when beer tax < $0.440 per gallon is greater than the probability mass for beer sales when beer tax >= $0.440 per gallon towards the right. A potential reason for this might be due too as beer tax is greater than $0.440 per gallon, people are less willing to purchase the same amount of beer as it is more expensive than when the tax is below $0.440 per gallon.

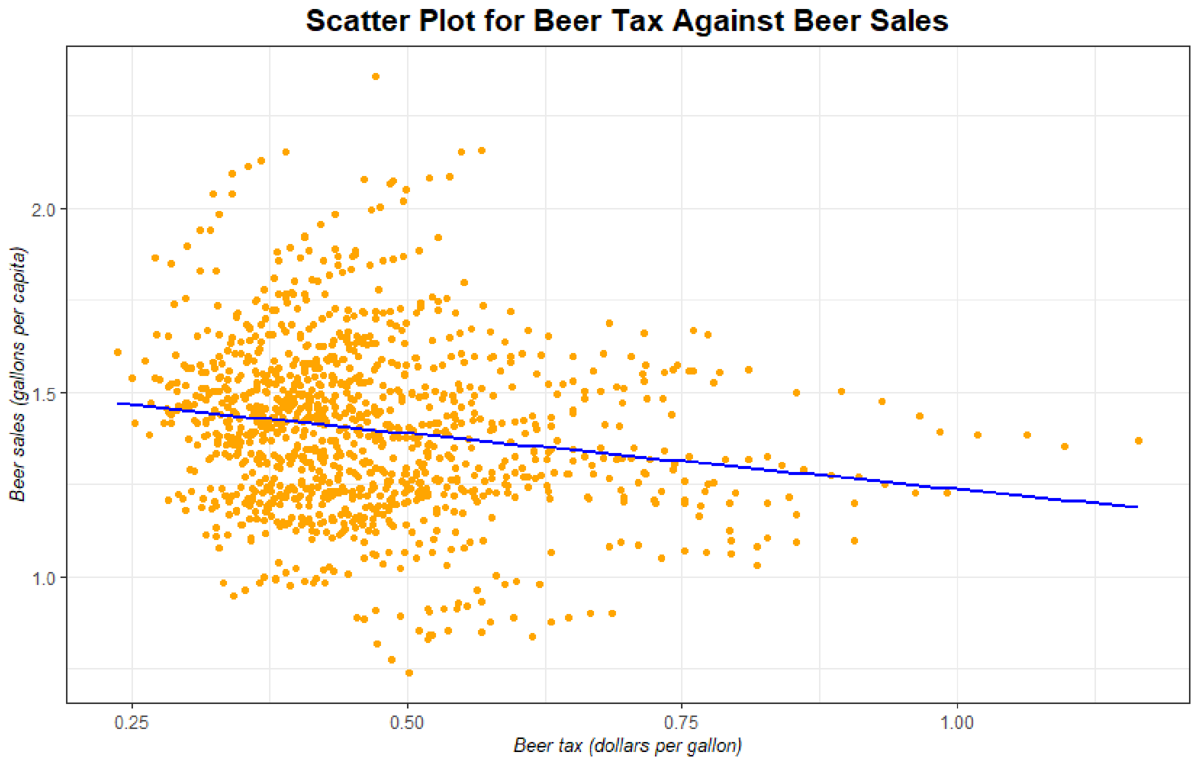
Question 4:

$\text{Difference of the means} = -0.061$
$95\%\ \text{CI:}\ [-0.087, -0.035]$

$\text{p\_value} = 5.941 \times 10^{-6}$

At the 5% significance level, we do not have enough evidence to support the null hypothesis that the mean of beer sales when beer tax is greater than $0.440 per gallon is the same as the mean of beer sales when beer tax is lower than $0.440 per gallon. Hence, we reject the null hypothesis and there's a statistically significantly difference. This is evident since there is a 4.23% decrease in beer sales when going from hightax = 0 to hightax = 1. Therefore, it may imply that as beer tax increase, we see a drop in beer sales.

Question 5:



Scatter Plot for Beer Tax Against Beer Sales

The scatter plot above shows the relationship between beer tax and beer sales. From the graph, it's visible that the relationship is a relatively weak negative relationship as the dots are crowded around the left side of the graph and there are less dots toward the right side.

Furthermore, the correlation coefficient between beer tax and beer sales is:

$$\text{corr}(\text{beer tax}, \text{beer sales}) = -0.173$$

As we can see, the correlation coefficient between the 2 variables is negative. This supports the findings from question 3 and 4 that as beer tax increase, we see a less amount of beer sales.

Question 6:

Regression 1		
Coefficients	Estimate	Standard error
Intercept	1.54148	0.02504
Beer tax	-0.30387	0.05151
Regression 2		
Coefficients	Estimate	Standard error
Intercept	1.4429630	0.0094745
Cigarette tax	-0.0012506	0.0001919

For regression 1, if there is a change of one-standard-deviation in beer tax, there will be a corresponding decrease of 0.304 * 0.130 = 0.039 gallons per capita in beer sales. Similarly, for regression 2, if there is a change of one-standard-deviation in cigarette tax, there will be a corresponding decrease of 0.0013 * 34.777 = 0.0345 gallons per capita in beer sales.

For regression 1:

$$\text{p\_value} = 4.804 \times 10^{-9}$$

$$95\% \text{ CI} = [-0.053, -0.026\,]$$

Hence, we reject the null hypothesis as p-value is smaller than the significance level of 5%. This implies that the predicted change in beer sales when there is a change in beer tax is statistically significantly different from 0.

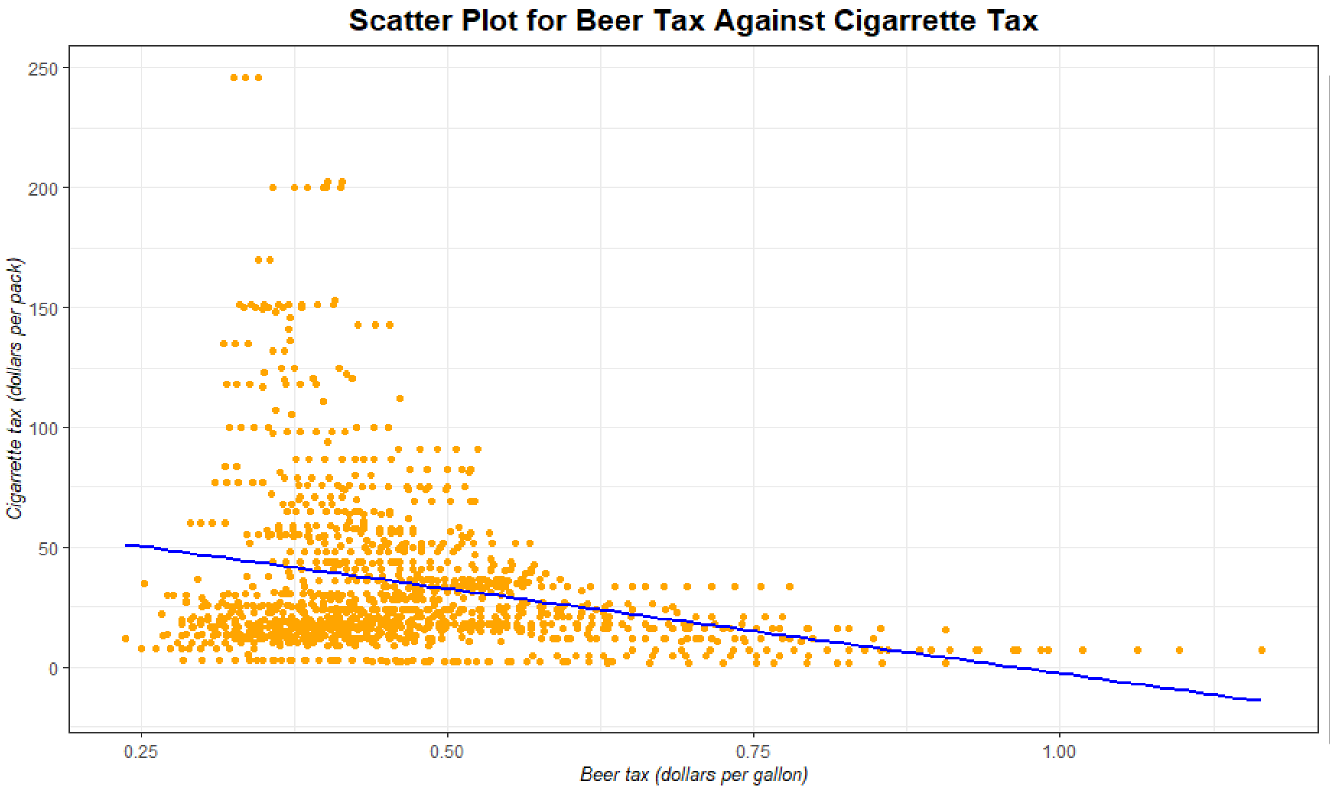
For regression 2:

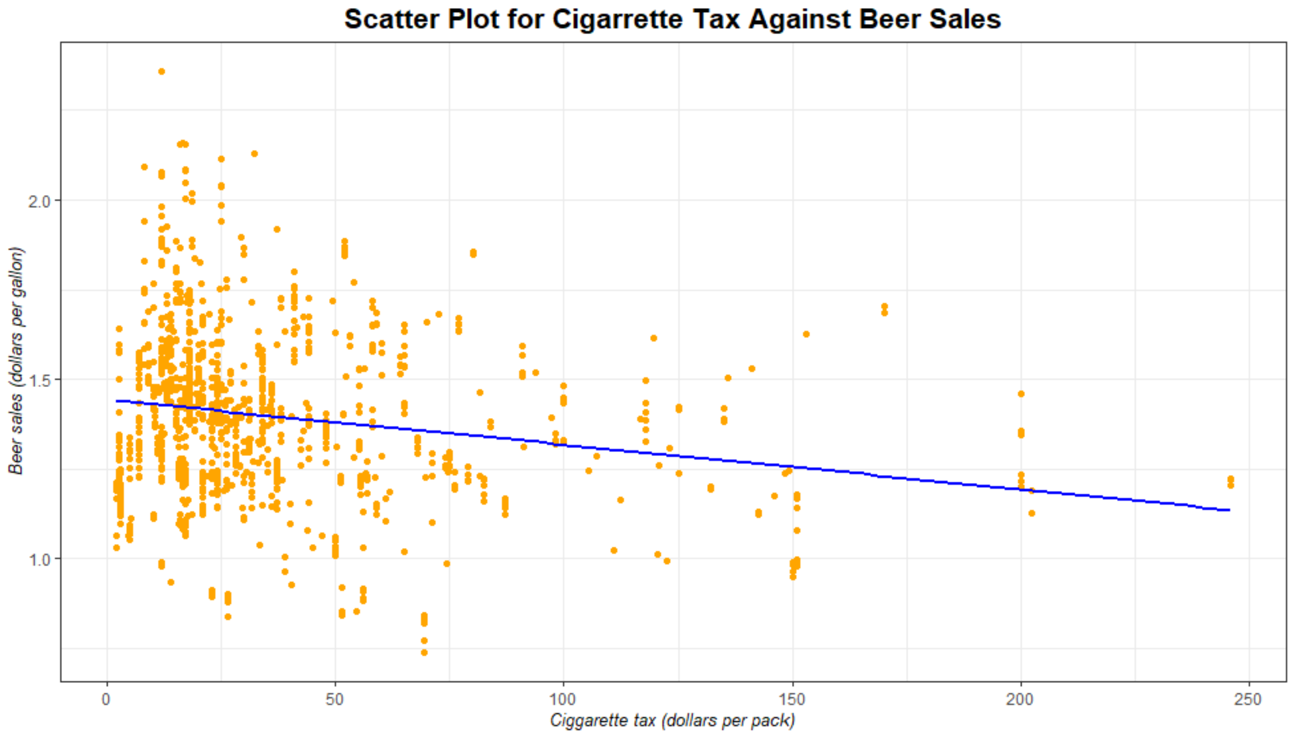
$$\text{p\_value} = 1.074 \times 10^{-10}$$

$$95\%\ CI = [-0.056, -0.0\,]$$

Hence, we reject the null hypothesis as p-value is smaller than the significance level of 5%. This implies that the predicted change in beer sales when there is a change in the cigarette tax is statistically significantly different from 0.

Question 7:





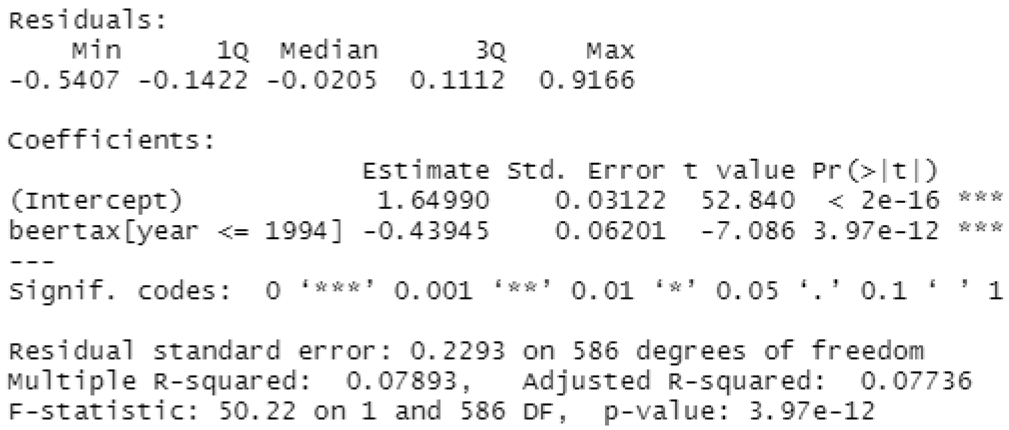
There is an increase in the coefficient estimate magnitude of beer sales in regression 3 compared with regression 1. This is because, in regression 1, cigarette tax was an omitted variable and was part of the residuals. However, from the scatter plot between beer tax and cigarette tax, we can see that there is a negative relationship which leads to our first assumptions of the OLS being inaccurate and the coefficient estimator on beer sales being higher than the actual value. Also, cigarette tax has a negative relationship with beer sales as well. Hence, the sign of the correlation between beer tax and the error terms is:

$$\text{sign}(\rho_{Xu}) = \text{sign}\big((-) \times (-)\big) = +$$

As a result, we see an increase in magnitude in the coefficient estimate of beer sales. The direction of the coefficient estimate stays the same.

Question 8:

Regression before and up to 1994:

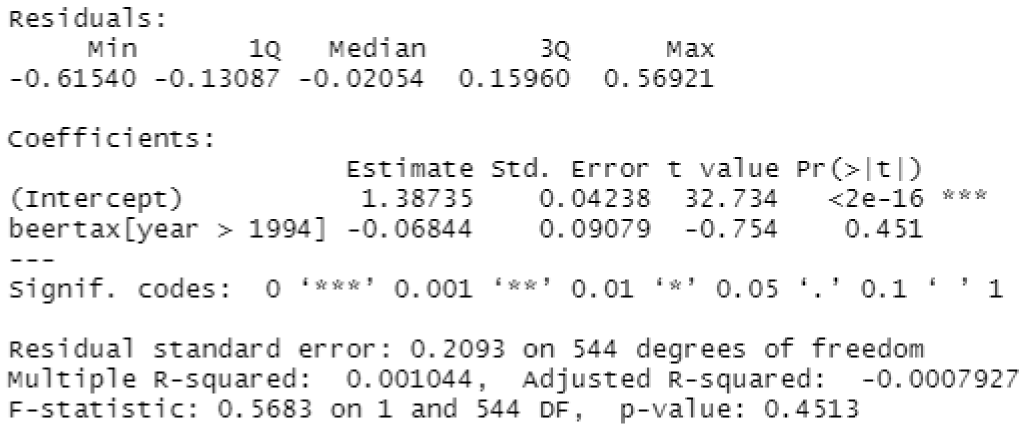


```
Residuals:
    Min      1Q  Median      3Q     Max 
-0.5407 -0.1422 -0.0205  0.1112  0.9166 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.64990    0.03122  52.840  < 2e-16 ***
beertax[year <= 1994] -0.43945    0.06201  -7.086 3.97e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2293 on 586 degrees of freedom
Multiple R-squared:  0.07893,   Adjusted R-squared:  0.07736 
F-statistic: 50.22 on 1 and 586 DF,  p-value: 3.97e-12
```

Regression after 1994:



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.61540 -0.13087 -0.02054  0.15960  0.56921 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.38735    0.04238  32.734   <2e-16 ***
beertax[year > 1994] -0.06844    0.09079  -0.754    0.451    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2093 on 544 degrees of freedom
Multiple R-squared:  0.001044,  Adjusted R-squared:  -0.0007927 
F-statistic: 0.5683 on 1 and 544 DF,  p-value: 0.4513
```

As we can see from the 2 figures, there is a drastic change in the coefficient estimator on beer sales between the 2 subsamples. At a 5% significance level, for the regression model before and up to 1994, we reject the null hypothesis that the coefficient estimator on beer sales is equal to 0 since its p-value is 3.97e-12 (less than 0.05). Hence, it is statistically significantly different from 0. In other words, before and up to 1994, there is a decent drop in beer sales if beer tax increases.

However, for the regression model after 1994, we failed to reject the null hypothesis since its p-value is 0.4513 (greater than 0.05). Thus, it is statistically significantly indifferent from 0. In other words, after 1994, there is little to no difference to beer sales even if beer tax increases.

Appendix

```
# Load the dataset into the environment and neccessary libraries
library('stargazer')
library('ggplot2')
library('dplyr')

# Question 1
# Dataset summary: contains annaual information on beer sales, beer taxes, and cigarret taxes from 1981 - 2007 in US (47 states)
beer_data = read.csv("D:/Semester 2 2022/ECOM20001/Assignment 1/as1_beer.csv")

# Report summary statistic
tax_st = stargazer(beer_data,
            summary.stat = c('n', 'mean', 'sd', 'median', 'min', 'max'),
            type = 'text', title = 'Descriptive Statistic for Beer Dataset',
            out = 'beer_sumstat.png')

# Question 2
# Compute sample means for each variable
beercons_mean = mean(beer_data$beercons)
beertax_mean = mean(beer_data$beertax)
cigtax_mean = mean(beer_data$cigtax)

# Compute sample standard error for each variable
beercons_se = sd(beer_data$beercons)/sqrt(length(beer_data$beercons))
beertax_se = sd(beer_data$beertax)/sqrt(length(beer_data$beertax))
cigtax_se = sd(beer_data$cigtax)/sqrt(length(beer_data$cigtax))

# Compute 95% Confidence Interval for each variable

# Beercons
```

```
beercons_95_low = beercons_mean - 1.96 * beercons_se

beercons_95_high = beercons_mean + 1.96 * beercons_se


# Beertax

beertax_95_low = beertax_mean - 1.96 * beertax_se

beertax_95_high = beertax_mean + 1.96 * beertax_se


# Cigtax

cigtax_95_low = cigtax_mean - 1.96 * cigtax_se

cigtax_95_high = cigtax_mean + 1.96 * cigtax_se



# Question 3


# adding a column of hightax straight into the dataset using dplyr.

beer_data2 = beer_data %>%

  mutate(hightax = if_else(beertax >= median(beertax),

                           1,

                           0))

beer_density = ggplot(beer_data2, aes(x = beercons, fill = as.character(hightax))) +

  geom_density(alpha = 0.25)


beer_density + labs(title = "Density Plot for Beer Sales",

                    x = "Beer sales (gallons per capita)",

                    y = "Density",

                    fill = "") +

            theme_bw() +

            scale_fill_manual(labels = c("Beer tax < $0.440", "Beer tax >= $0.440"), values = c("Blue", "Red")) +

            theme(plot.title = element_text(face = "bold", size = 16, hjust = 0.5),

                  axis.title = element_text(face = "italic", size = 10))
```

```
beer_density

# Question 4

# Compute mean for beercons if hightax = 1 and hightax = 0
beercons_ht1_mean = mean(beer_data2$beercons[beer_data2$hightax == 1])
beercons_ht0_mean = mean(beer_data2$beercons[beer_data2$hightax == 0])

# Compute standard deviation and standard errors
beercons_ht1_sd = sd(beer_data2$beercons[beer_data2$hightax == 1])
beercons_ht0_sd = sd(beer_data2$beercons[beer_data2$hightax == 0])

beercons_ht1_se = beercons_ht1_sd/sqrt(length(beer_data2$beercons[beer_data2$hightax == 1]))
beercons_ht0_se = beercons_ht0_sd/sqrt(length(beer_data2$beercons[beer_data2$hightax == 0]))

mean_diff = beercons_ht1_mean - beercons_ht0_mean
se_diff = sqrt((beercons_ht1_se**2 + beercons_ht0_se**2))

# Conduct hypothesis testing:
t_stat = (mean_diff - 0)/se_diff
p_value = 2 * pnorm(-abs(t_stat))

# 95% CI for mean diff
mean_diff_upper = mean_diff + 1.96 * se_diff
mean_diff_lower = mean_diff - 1.96 * se_diff

diff_percentage = (abs(mean_diff)/beercons_ht0_mean) * 100

# There is a decrease of approximately 4% when hightax went from 0 to 1.
# Question 5

#scatter plot of beertax against beercons
beer_scatter = ggplot(beer_data2, aes(beertax, beercons)) + geom_point(colour = 'orange')
beer_scatter + labs(title = "Scatter Plot for Beer Tax Against Beer Sales",
             x = "Beer tax (dollars per gallon)",
             y = "Beer sales (gallons per capita)") +
           theme_bw() +
           theme(plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
               axis.title = element_text(face = "italic", size = 10)) +
           geom_smooth(method = lm, se = FALSE, colour = "blue")


cor_beertax_cons = cor(beer_data2$beertax, beer_data2$beercons)

# Question 6

# Single linear regression for beertax and beercons
reg_1 = lm(beercons~beertax, beer_data2)
summary(reg_1)

# Single linear regressions for cigtax and beercons,
reg_2 = lm(beercons~cigtax, beer_data2)
summary(reg_2)

# standard deviation for beertax and cigtax
beertax_sd = sd(beer_data2$beertax)
cigtax_sd = sd(beer_data2$cigtax)
```

```
# Increase of 1 standard-deviation in beertax
change_1 = -0.30387 * beertax_sd
change_1

# increase of 1-standard-deviation in cigtax
change_2 = -0.0012506 * cigtax_sd
change_2

# Question 7

# scatter plot for beercons and cigtax

cons_cig_scatter = ggplot(beer_data2, aes(cigtax, beercons)) + geom_point(colour = 'orange')
cons_cig_scatter = cons_cig_scatter + labs(title = "Scatter Plot for Cigarrette Tax Against Beer Sales",
                    y = "Beer sales (dollars per gallon)",
                    x = "Ciggarette tax (dollars per pack)") +
                theme_bw() +
                theme(plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
                      axis.title = element_text(face = "italic", size = 10)) +
                geom_smooth(method = lm, se = FALSE, colour = "blue")
cons_cig_scatter

# scatter plot for beertax and cigtax

beer_cig_scatter = ggplot(beer_data2, aes(beertax, cigtax)) + geom_point(colour = "orange")
beer_cig_scatter = beer_cig_scatter + labs(title = "Scatter Plot for Beer Tax Against Cigarrette Tax",
                    x = "Beer tax (dollars per gallon)",
                    y = "Cigarrette tax (dollars per pack)") +
                theme_bw() +
            theme(plot.title = element_text(face = "bold", size = 16, hjust = 0.5),
                axis.title = element_text(face = "italic", size = 10)) +
            geom_smooth(method = lm, se = FALSE, colour = "blue")

beer_cig_scatter


# QUestion 8
reg_before_1994 = lm(beercons[year <= 1994]~beertax[year <= 1994], beer_data2)
summary(reg_before_1994)

reg_after_1994 = lm(beercons[year > 1994]~beertax[year > 1994], beer_data2)
summary(reg_after_1994)

# Hypothesis testing for difference in beertax coefficients

coef_diff = -0.43945 - (-0.06844)
se_diff2 = sqrt(0.06201**2 + 0.09079**2)

t_stat2 = (coef_diff - 0)/se_diff2
p_value2 = 2*pnorm(-abs(t_stat2)) # = 0.000739545

# since p value is less than 0.05, we reject the null hypothesis and there is statistically significant difference
# between the coefficients of the two sets.
```