Homework 3-1

Yaseen Bhat

Link to Github

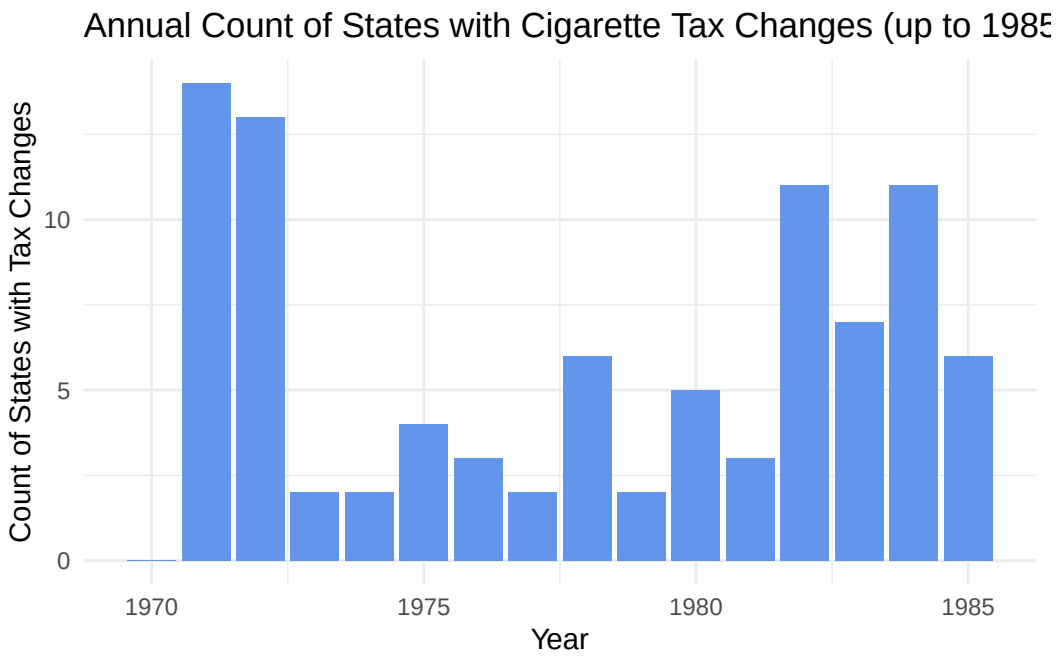
Summarize The Data

1. Present a bar graph showing the proportion of states with a change in their cigarette tax in each year from 1970 to 1985.

```
tax_changes_by_year <- cig.data %>%
  filter(Year <= 1985) %>%
  group_by(Year) %>%
  summarise(TaxChangeCount = sum(tax_change_d, na.rm = TRUE))

  tax_change_plot <- ggplot(tax_changes_by_year, aes(x = Year, y = TaxChangeCount)) +
  geom_col(fill = "cornflowerblue") +
  labs(title = "Annual Count of States with Cigarette Tax Changes (up to 1985)",
       x = "Year",
       y = "Count of States with Tax Changes") +
  theme_minimal()


print(tax_change_plot)
```



2

. Plot on a single graph the average tax (in 2012 dollars) on cigarettes and the average price of a pack of cigarettes from 1970 to 2018.

```
data.2018 <- cig.data %>%
filter(Year <= 2018)%>%
group_by(Year)%>%
summarise(mean_price = mean(price_cpi_2022, na.rm = TRUE), mean_tax = mean(total_tax_cpi_202

data.2018_plot <- ggplot(data.2018, aes(x = Year))+
geom_line(aes(y = mean_price), color = "black")+
geom_line(aes(y = mean_tax), color = "yellow")
labs(title = "Average Tax and Price of Cigarettes from 1970 to 2018)",
       x = "Year", y = "Value in 2012 dollars") +
  theme_classic()
```

```
NULL
```

```
print(data.2018)
```

```
# A tibble: 49 x 3
    Year mean_price mean_tax
   <dbl>      <dbl>    <dbl>
 1  1970       2.25    1.06
 2  1971       2.24    1.08
 3  1972       2.19    1.07
 4  1973       2.14    1.02
 5  1974       2.05    0.920
 6  1975       2.03    0.856
 7  1976       1.97    0.815
 8  1977       2.06    0.777
 9  1978       1.99    0.723
10  1979       1.89    0.656
# i 39 more rows
```

3

. Identify the 5 states with the highest increases in cigarette prices (in dollars) over the time period. Plot the average number of packs sold per capita for those states from 1970 to 2018.

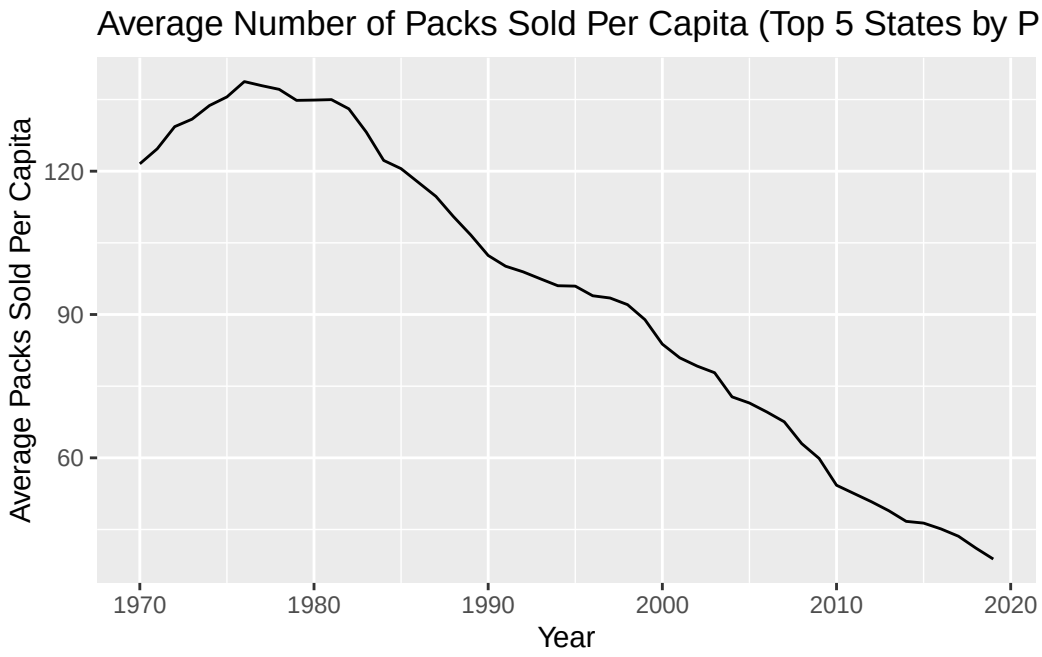
```
price_changes <- cig.data %>%
  filter(Year %in% c(1970, 2018)) %>%
  spread(key = Year, value = price_cpi_2022) %>%
  mutate(PriceIncrease = `2018` - `1970`) %>%
  select(state, PriceIncrease) %>%
  arrange(desc(PriceIncrease)) %>%
  slice_head(n = 5)

top_states <- price_changes$state

# calc n plot
avg_sales_top_states <- cig.data %>%
  filter(state %in% top_states) %>%
  group_by(Year) %>%
  summarise(MeanSales = mean(sales_per_capita, na.rm = TRUE))

avg_sales_plot <- ggplot(avg_sales_top_states, aes(x = Year, y = MeanSales)) +
  geom_line() +
  labs(title = "Average Number of Packs Sold Per Capita (Top 5 States by Price Increase)",
       x = "Year",
       y = "Average Packs Sold Per Capita")


print(avg_sales_plot)
```



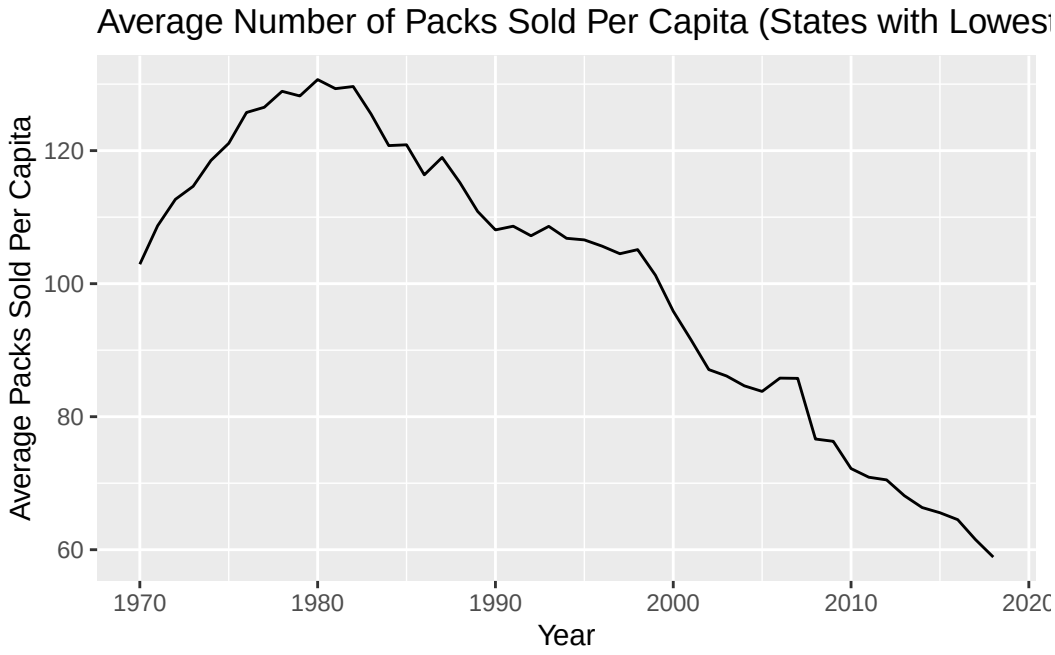
4. Identify the 5 states with the lowest increases in cigarette prices (in dollars) over the time period. Plot the average number of packs sold per capita for those states from 1970 to 2018.

```
bottom_states <- c("Missouri", "Tennessee", "North Dakota", "Alabama", "Georgia")

# calc n plot
avg_sales_bottom_states <- cig.data %>%
  filter(state %in% bottom_states, Year <= 2018) %>%
  group_by(Year) %>%
  summarise(MeanSales = mean(sales_per_capita, na.rm = TRUE))

avg_sales_bottom_plot <- ggplot(avg_sales_bottom_states, aes(x = Year, y = MeanSales)) +
  geom_line() +
  labs(title = "Average Number of Packs Sold Per Capita (States with Lowest Price Increase)"
       x = "Year",
       y = "Average Packs Sold Per Capita")

# Display the plot
print(avg_sales_bottom_plot)
```


Average Number of Packs Sold Per Capita (States with Lowes
Average Packs Sold Per Capita
120
100
80
60
1970
1980
1990
2000
2010
2020
Year

5

. Compare the trends in sales from the 5 states with the highest price increases to those with the lowest price increases.

Both sets of groups showed a decreasing trend of cig sales over time. But the states with the higher price increases led to lesser sales than the ones with the lower increases.

Estimate ATEs

Now let's work on estimating a demand curve for cigarettes. Specifically, we're going to estimate the price elasticity of demand for cigarettes. When explaining your findings, try to limit your discussion just to a couple of sentences.

6. Focusing only on the time period from 1970 to 1990, regress log sales on log prices to estimate the price elasticity of demand over that period. Interpret your results.

```
price_elasticity_model <- lm(ln_sales ~ ln_price_2012, data = filter(cig.data, Year >= 1970 &

model_summary <- summary(price_elasticity_model)
print(model_summary)
```

```
Call:
lm(formula = ln_sales ~ ln_price_2012, data = filter(cig.data,
    Year >= 1970 & Year <= 1990))

Residuals:
     Min       1Q   Median       3Q      Max
-0.68335 -0.08598 -0.00284  0.08778  0.83516

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.42738    0.02975   182.4   <2e-16 ***
ln_price_2012 -0.80944    0.03837   -21.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1894 on 1069 degrees of freedom
Multiple R-squared:  0.294, Adjusted R-squared:  0.2933
F-statistic: 445.1 on 1 and 1069 DF,  p-value: < 2.2e-16
```

The coefficient is negative, which means as price goes up, sales go down and vice-versa.

7. Again limiting to 1970 to 1990, regress log sales on log prices using the total (federal and state) cigarette tax (in dollars) as an instrument for log prices. Interpret your results and compare your estimates to those without an instrument. Are they different? If so, why?

```r
regdata <- cig.data %>%
  filter(Year >= 1970 & Year <= 1990)
summary(feols(ln_sales ~ 1 | ln_price_2012 ~ ln_tax_2012,
              data=regdata))
```

```
TSLS estimation, Dep. Var.: ln_sales, Endo.: ln_price_2012, Instr.: ln_tax_2012
Second stage: Dep. Var.: ln_sales
Observations: 1,071
Standard-errors: IID
                  Estimate Std. Error  t value  Pr(>|t|)
(Intercept)       5.416797   0.054494  99.4017 < 2.2e-16 ***
fit_ln_price_2012 -0.795524   0.071235 -11.1676 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.189226   Adj. R2: 0.293235
F-test (1st stage), ln_price_2012: stat = 436.8     , p < 2.2e-16 , on 1 and 1,069 DoF.
                      Wu-Hausman: stat =   0.053709, p = 0.816775, on 1 and 1,068 DoF.
```

The significant negative coefficient for ln_price_2012 confirms that higher prices lead to lower sales. It doesn't seem like there's a significant difference with the estimates without the instrument (in this case the ln_tax_2012)

8. Show the first stage and reduced-form results from the instrument.

```r
#q8
price_elasticity_model <- cig.data %>%
  filter(Year >= 1970 & Year <= 1990)

first_stage_model <- lm(ln_price_2012 ~ ln_tax_2012, data=price_elasticity_model)

price_elasticity_model$pricehat <- predict(first_stage_model)

second_stage_model <- lm(ln_sales ~ pricehat, data=price_elasticity_model)

summary(first_stage_model)
```

```
Call:
lm(formula = ln_price_2012 ~ ln_tax_2012, data = price_elasticity_model)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-0.23046 -0.09207 -0.02919  0.08019  0.48675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.839646   0.005421   154.9   <2e-16 ***
ln_tax_2012 0.260060   0.012443    20.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1272 on 1069 degrees of freedom
Multiple R-squared:  0.2901,    Adjusted R-squared:  0.2894
F-statistic: 436.8 on 1 and 1069 DF,  p-value: < 2.2e-16
```

```
summary(second_stage_model)
```

```

Call:
lm(formula = ln_sales ~ pricehat, data = price_elasticity_model)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86239 -0.09798  0.00549  0.09359  0.95094

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.41680    0.06212  87.196   <2e-16 ***
pricehat    -0.79552    0.08121  -9.796   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2159 on 1069 degrees of freedom
Multiple R-squared:  0.08238,   Adjusted R-squared:  0.08152
F-statistic: 95.97 on 1 and 1069 DF,  p-value: < 2.2e-16
```

```
reduced_form_model <- lm(ln_sales ~ ln_tax_2012, data=price_elasticity_model)
summary(reduced_form_model)
```

```

Call:
lm(formula = ln_sales ~ ln_tax_2012, data = price_elasticity_model)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.86239 -0.09798  0.00549  0.09359  0.95094

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.748839   0.009202 516.092   <2e-16 ***
ln_tax_2012 -0.206884   0.021119  -9.796   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2159 on 1069 degrees of freedom
Multiple R-squared:  0.08238,   Adjusted R-squared:  0.08152
F-statistic: 95.97 on 1 and 1069 DF,  p-value: < 2.2e-16
```

The coefficient for the second stage is -0.207 which is less an effect than the first stage result of -0.412. This is unexpected and may be a result of error as one might expect the decrease to be larger when acounting for the endogeneity of price.

9

. Repeat questions 1-3 focusing on the period from 1991 to 2015.

```
#q9
regdata2 <- cig.data %>%
filter(Year >= 1991 & Year <= 2015)

first_stage_model2 <- lm(ln_sales ~ ln_price_2012, data = regdata2)
summary(first_stage_model2)
```

```
Call:
lm(formula = ln_sales ~ ln_price_2012, data = regdata2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92230 -0.17004  0.00664  0.17869  1.10282

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.65996    0.03638  155.56   <2e-16 ***
ln_price_2012 -0.99681    0.02469  -40.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.296 on 1273 degrees of freedom
Multiple R-squared:  0.5614,    Adjusted R-squared:  0.5611
F-statistic:  1630 on 1 and 1273 DF,  p-value: < 2.2e-16
```

```
summary(feols(ln_sales ~ 1 | ln_tax_2012,
              data= regdata2))
```

```
OLS estimation, Dep. Var.: ln_sales
Observations: 1,275
Fixed-effects: ln_tax_2012: 1,024
RMSE: 0.112223   Adj. R2: 0.679548
```

```
first_stage2 <- lm(ln_price_2012 ~ ln_tax_2012, data = regdata2)
summary(first_stage2)
```

```
Call:
lm(formula = ln_price_2012 ~ ln_tax_2012, data = regdata2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36750 -0.09020  0.00725  0.08241  0.45045

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.315073   0.004386  299.84   <2e-16 ***
ln_tax_2012 0.513550   0.006922   74.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1456 on 1273 degrees of freedom
Multiple R-squared:  0.8121,    Adjusted R-squared:  0.812
F-statistic:  5504 on 1 and 1273 DF,  p-value: < 2.2e-16
```

```r
reduced_form_model2 <- lm(ln_sales ~ ln_tax_2012, data = regdata2)
summary(reduced_form_model2)
```

```
Call:
lm(formula = ln_sales ~ ln_tax_2012, data = regdata2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90878 -0.15465  0.01119  0.15334  1.16925

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.36742    0.00844  517.50   <2e-16 ***
ln_tax_2012 -0.59063    0.01332  -44.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2802 on 1273 degrees of freedom
Multiple R-squared:  0.607, Adjusted R-squared:  0.6067
F-statistic:  1966 on 1 and 1273 DF,  p-value: < 2.2e-16
```

I made a mistake in this code, spent so much time trying to figure it out but i kept getting

new errors. Will redo it in smaller chunks for submission 2. Tried to reset it and do it smaller chunks but it was something i was doing wrong clearly.

10. Compare your elasticity estimates from 1970-1990 versus those from 1991-2015. Are they different? If so, why?.

I'm assuming that the 1970 one will be larger? As my results are skewed for number 9, I am unsure right now, however juding by my results, this is still kind of the case. As there is a strong negative relationship between tax and sales per capita.