HW3

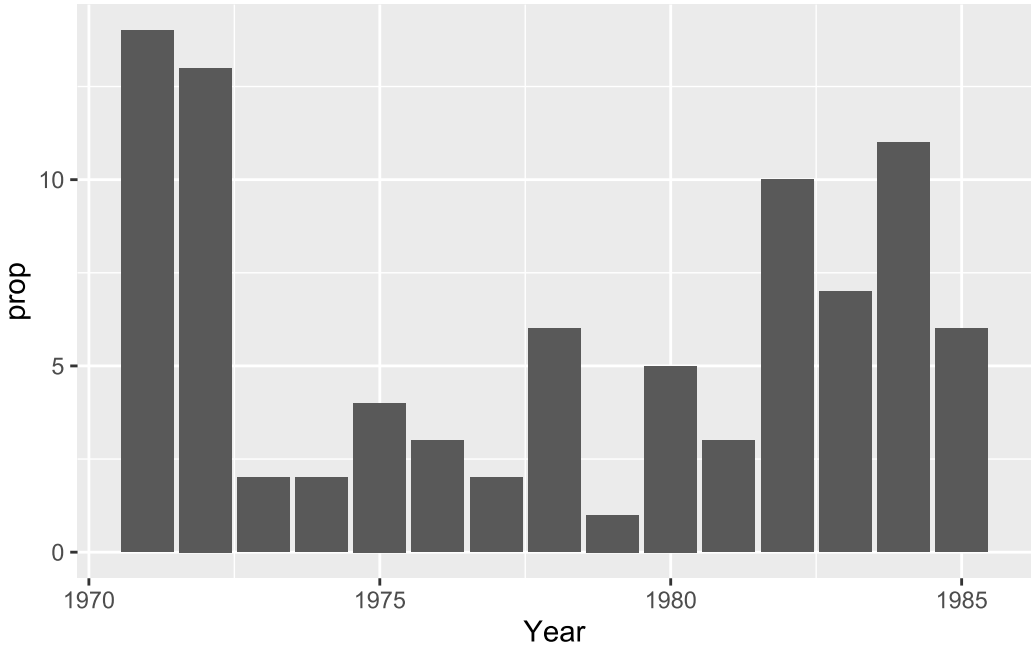
Miracle Ephraim

Question 1

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

```
#| echo: false
ggplot(data=prop_tax, aes(Year,prop)) +
    geom_col()
```

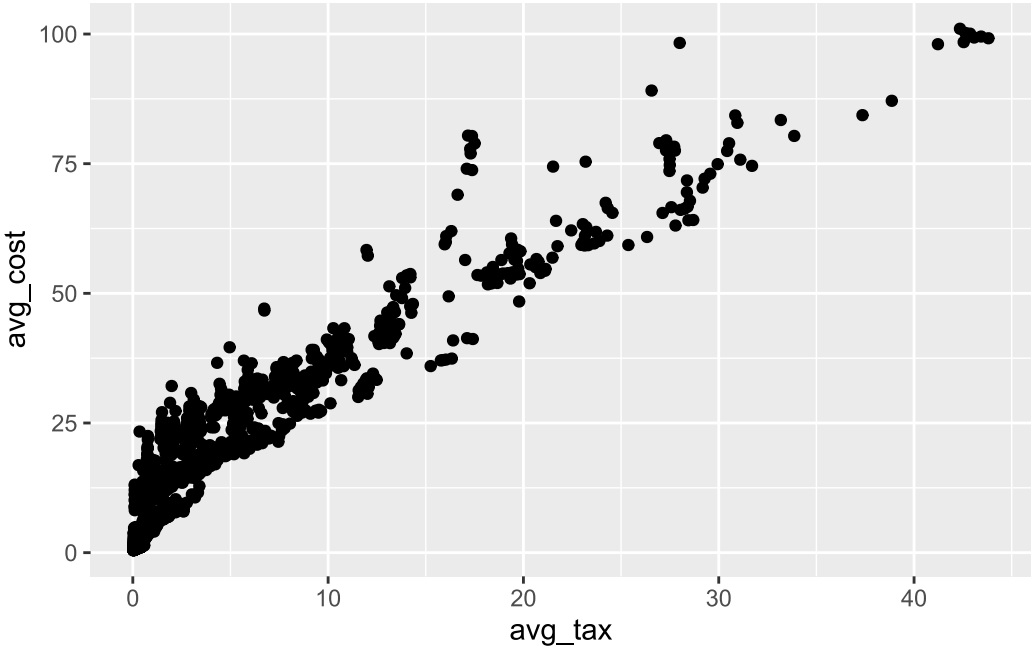
```
Warning: Removed 1 row containing missing values or values outside the scale
range
(`geom_col()`).
```



Question 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.

```
#| echo: false
ggplot(data = avgs, aes(avg_tax, avg_cost)) +
    geom_point()
```



Question 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.

```
#| echo: false

top5 <- df %>%
    filter(Year == 1970 | Year == 2018) %>%
    group_by(state) %>%
    summarise(diff = cost_per_pack - lag(cost_per_pack)) %>%
    filter(!is.na(diff)) %>%
    arrange(desc(diff))
```

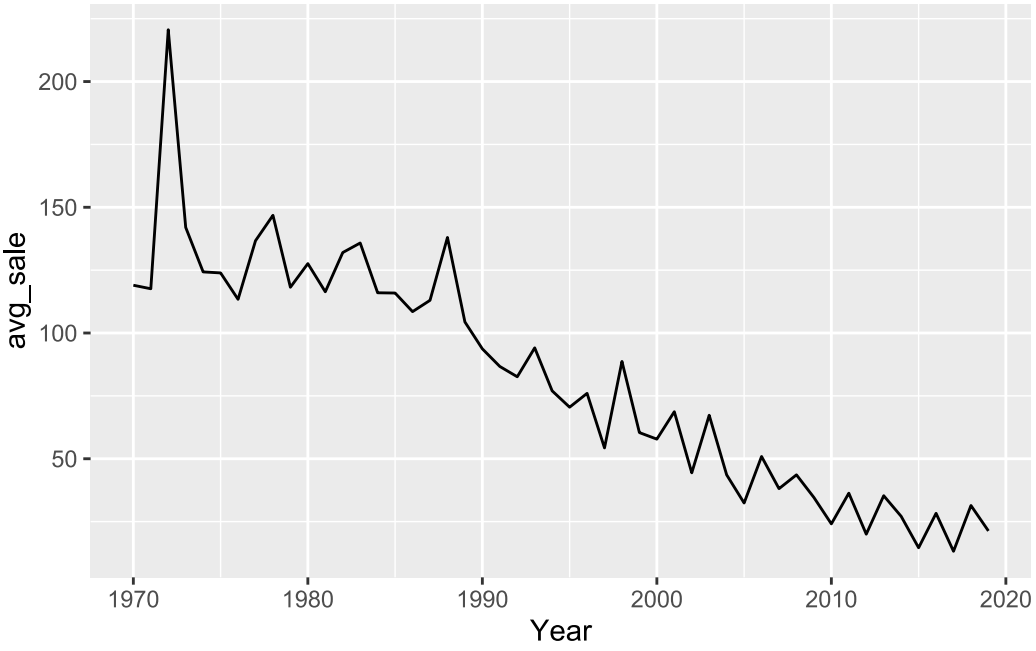
```
Warning: Returning more (or less) than 1 row per `summarise()` group was
deprecated in
dplyr 1.1.0.
i Please use `reframe()` instead.
i When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.
```

```
`summarise()` has grouped output by 'state'. You can override using the
`.groups` argument.
```

```
head(top5, 5)
```

```
# A tibble: 5 × 2
# Groups:   state [5]
  state                diff
  <chr>               <dbl>
1 New York             9.93
2 District of Columbia 9.54
3 Connecticut          9.30
4 Rhode Island         9.20
5 Massachusetts        9.18
```

```
ggplot(change, aes(Year, avg_sale)) +
    geom_line()
```



Question 4

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

```

```

```
#| echo: false

low5 <- df %>%
    filter(Year == 1970 | Year == 2018) %>%
    group_by(state) %>%
    summarise(diff = cost_per_pack - lag(cost_per_pack)) %>%
    filter(!is.na(diff)) %>%
    arrange(diff)
```

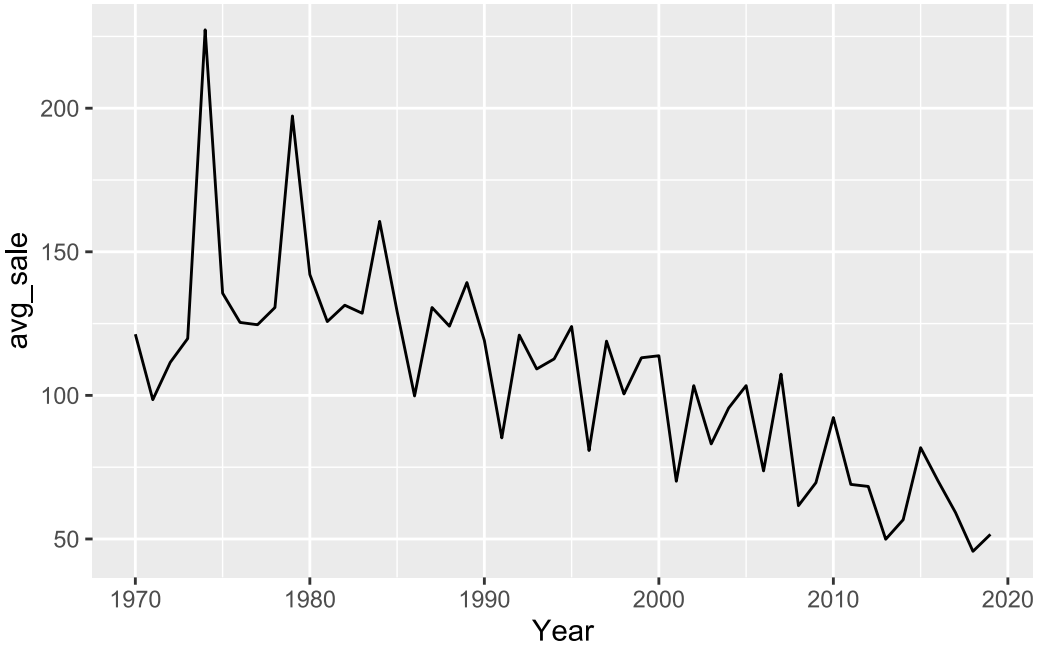
```
Warning: Returning more (or less) than 1 row per `summarise()` group was
deprecated in
dplyr 1.1.0.
ℹ Please use `reframe()` instead.
ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.
```

```
`summarise()` has grouped output by 'state'. You can override using the
`.groups` argument.
```

```
head(low5, 5)
```

```
# A tibble: 5 × 2
# Groups:   state [5]
  state          diff
  <chr>         <dbl>
1 Missouri       4.59
2 North Dakota   4.85
3 Tennessee      4.86
4 Georgia        4.87
5 North Carolina 4.88
```

```
ggplot(change2, aes(Year, avg_sale)) +
    geom_line()
```



Question 5

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

Amongst states with the highest prices, there has been a overall decrease in the average sales of cigarettes in the last few decades. While states with the lowest prices also have seen an overall decrease, sales following significant moments where there significant drops in price are much higher than similar moments in the top 5 states.

Question 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.

```
summary(ols1)
```

```
Call:
lm(formula = log_sales ~ log_cost, data = (df %>% filter(Year >=
    1970 & Year <= 1990)))

Residuals:
     Min       1Q   Median       3Q      Max
-0.77629 -0.09967 -0.00787  0.09969  0.78423
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.750402   0.008116   585.3   <2e-16 ***
log_cost    -0.171540   0.013829   -12.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2107 on 1069 degrees of freedom
Multiple R-squared:  0.1258,    Adjusted R-squared:  0.125
F-statistic: 153.9 on 1 and 1069 DF,  p-value: < 2.2e-16
```

On average, quantity demanded decreases by 17% for every 1% increase in price.

Question 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?

```
summary(step2)
```

```
Call:
lm(formula = log_sales ~ pricehat, data = (df %>% filter(Year >=
    1970 & Year <= 1990)))

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)  4.99111    0.01947 256.365   <2e-16 ***
pricehat     0.50237    0.05128   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
```

With a coefficient of 0.5, we can say price of cigarettes is inelastic, with a 1% increase in price leading to only a 0.5% change in quantity demanded. These are quite different from the OLS estimates of elasticity, likely due to the endogeneity present in the OLS model.

Question 8

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

```
summary(step1)
```

```
Call:
lm(formula = log_cost ~ log_total_tax, data = (df %>% filter(Year >=
    1970 & Year <= 1990)))

Residuals:
    Min      1Q  Median      3Q     Max
-1.0060 -0.3359 -0.1095  0.3779  1.0872

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.48152    0.01904 -25.296   <2e-16 ***
log_total_tax -0.41181    0.04381  -9.399   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4479 on 1069 degrees of freedom
Multiple R-squared:  0.07634,   Adjusted R-squared:  0.07547
F-statistic: 88.35 on 1 and 1069 DF,  p-value: < 2.2e-16
```

```
summary(step2)
```

```
Call:
lm(formula = log_sales ~ pricehat, data = (df %>% filter(Year >=
    1970 & Year <= 1990)))

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)  4.99111    0.01947 256.365   <2e-16 ***
pricehat     0.50237    0.05128   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
```

Question 9

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

OLS estimates

```
summary(ols2)
```

```

Call:
lm(formula = log_sales ~ log_cost, data = (df %>% filter(Year >=
    1991 & Year <= 2015)))

Residuals:
    Min      1Q  Median      3Q     Max
-0.9375 -0.1781  0.0013  0.1860  1.1433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.03949    0.02291  219.93   <2e-16 ***
log_cost    -0.66563    0.01747  -38.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3056 on 1273 degrees of freedom
Multiple R-squared:  0.5327,    Adjusted R-squared:  0.5323
F-statistic:  1451 on 1 and 1273 DF,  p-value: < 2.2e-16
```

As price increases by 1%, quantity demanded decreases by 66%.

IV estimates

```
summary(step4)
```

```

Call:
lm(formula = log_sales ~ pricehat, data = (df %>% filter(Year >=
    1991 & Year <= 2015)))

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)  5.21890    0.02365  220.69   <2e-16 ***
pricehat    -0.81311    0.01834  -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
```

With a coefficient of −0.8, we can say price of cigarettes is inelastic, with a 1% increase in price leading to only a 0.8% change in quantity demanded. These are different from the OLS estimates of elasticity as well, but not as stark of a contrast from the OLS estimates.

2SLS estimates

```
summary(step3)
```

```

Call:
lm(formula = log_cost ~ log_total_tax, data = (df %>% filter(Year >=
    1991 & Year <= 2015)))

Residuals:
     Min       1Q   Median       3Q      Max
-0.57946 -0.17145  0.02125  0.16675  0.66985

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.045904   0.007185  145.57   <2e-16 ***
log_total_tax 0.726380   0.011328   64.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2383 on 1273 degrees of freedom
Multiple R-squared:  0.7636,    Adjusted R-squared:  0.7634
F-statistic:  4111 on 1 and 1273 DF,  p-value: < 2.2e-16
```

```
summary(step4)
```

```

Call:
lm(formula = log_sales ~ pricehat, data = (df %>% filter(Year >=
    1991 & Year <= 2015)))
```

```
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)  5.21890    0.02365  220.69   <2e-16 ***
pricehat    -0.81311    0.01834  -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
```

Question 10

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

The estimates from each 20-year period are different from eachother, likely due to changes in attitudes around smoking and the increasing taxes placed on the product.