DS102-10

Lesson 10 Final Project

Heather Walker
2022-08-22

PROJECT INSTRUCTIONS

- Demonstrate the skills and knowledge you've gained to explore one last data set using R.
- Document your exploration in an R script file
- Create slide presentation that includes:
  - R code used in each step
  - Plots created with R code
  - Description of the information you found in the data and plots
  - Describe your interpretation of the plots and information provided by R.

TABLE OF CONTENTS

01 CREATE A BOX PLOT

02 FIND THE MEDIAN

03 CREATE A SCATTER PLOT

04 ADJUST FOR INFLATION

05 COMPARE 1985 to 1995

05+ GRAPHING DATA FOR DEP. T-TEST

TABLE OF CONTENTS (continued)

06 INDEPENDENT ANALYSIS:
Relationship between income and cigarette pack sales per capita.

07 INDEPENDENT ANALYSIS:
Packs sold per capita in the state of TN vs. the rest of the United States.

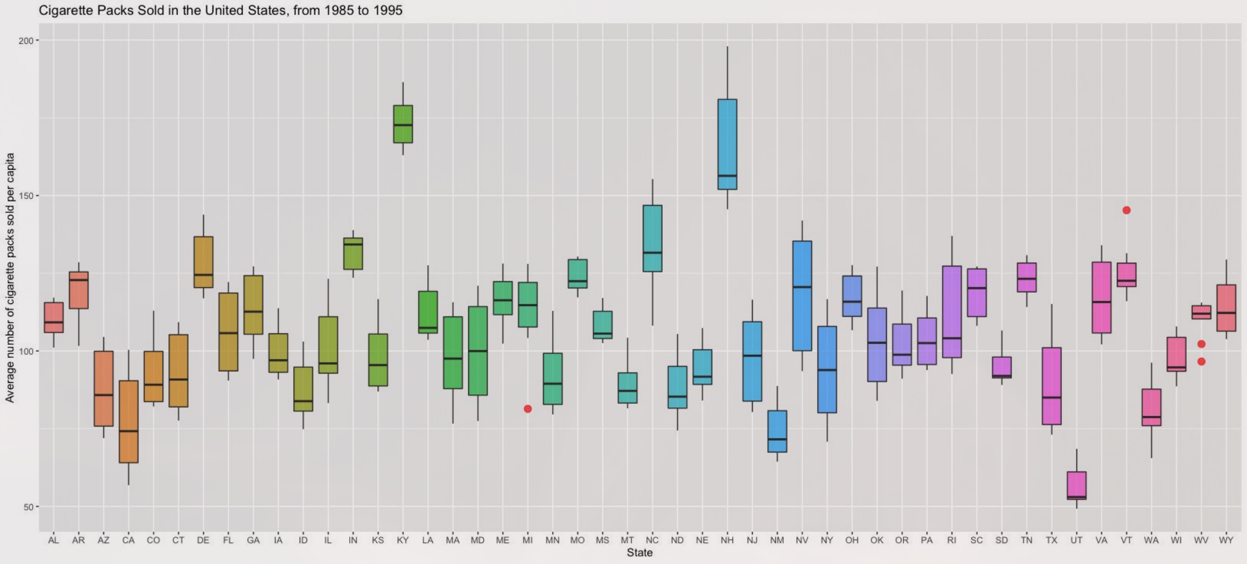
01

CREATE A BOX PLOT

Create a box plot of the average number of packs per capita by state.

CREATE A BOX PLOT

```
ggplot(Cigarette, aes(x=state,
y=packpc, fill=state)) +
  geom_boxplot(

  # additional color settings
  alpha=0.7,

  # custom outliers
  outlier.color="red",
  outlier.fill="red",
  outlier.size=3

  ) +
  theme(legend.position="none") +
  xlab("State") + ylab("Average
number of cigarette packs sold
per capita") +
  ggtitle("Cigarette Packs Sold in
the United States, from 1985 to
1995")
```



INSIGHTS FROM THE DATA

○ Which states have the highest number of packs?

These states are well above the rest of the group:
(In alphabetical order)
- Kentucky (KY)
- North Carolina (NC)
- New Hampshire (NH)

○ Which states have the lowest number of packs?

This state is well below the rest of the group:
- Utah (UT) – perhaps due to the large population of Mormons?

Other states that are lowest, but not as low:
- New Mexico (NM)
- Washington (WA)

02

FIND THE MEDIAN

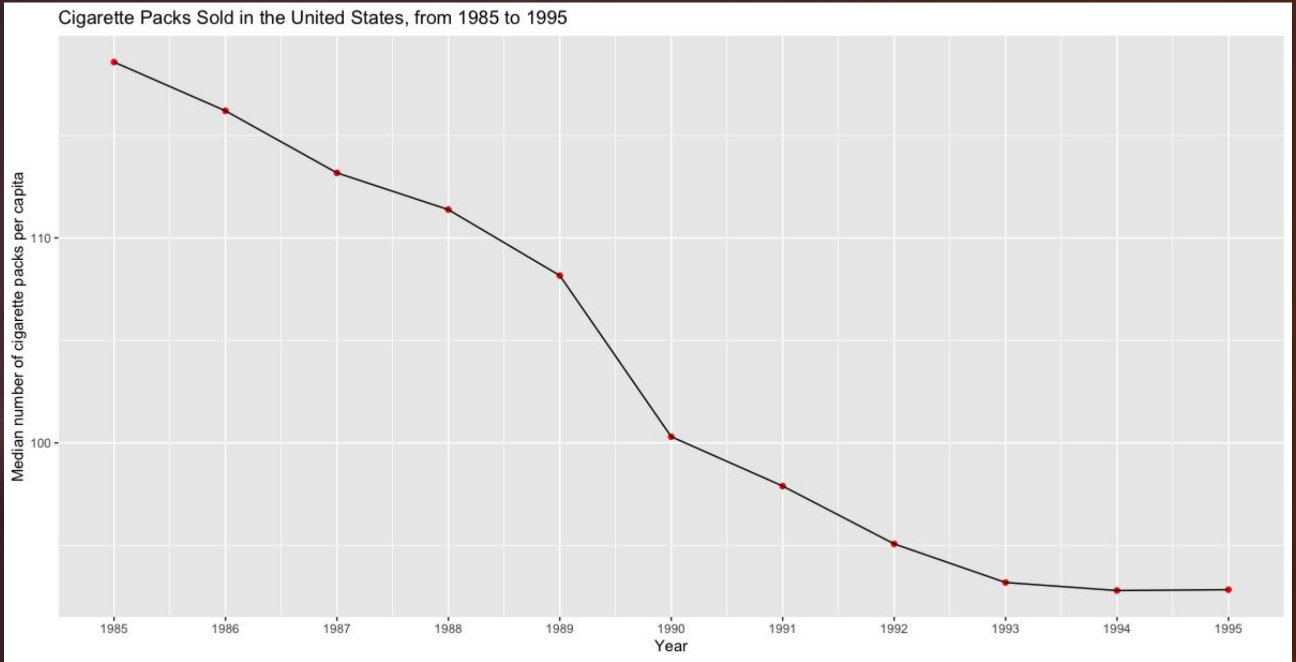
Find the median over all the states of the number of packs per capita per year.

FIND THE MEDIAN

```
# combine summarize() with group_by() and store in variable
medianPackPC <- Cigarette %>% group_by(year) %>% summarize(yearMedian = median(packpc))

# look at medianPackPC
View(medianPackPC)
medianPackPC

# Seeing the min and max for y-axis
range(medianPackPC$yearMedian)

# plot medianPackPC - line plot
ggplot(medianPackPC, aes(x=year, y=yearMedian)) +
  geom_point(color="red") +
  geom_line() +
  scale_x_continuous(breaks = seq(1985,1995)) +
  xlab("Year") + ylab("Median number of cigarette packs per capita") +
  ggtitle("Cigarette Packs Sold in the United States, from 1985 to 1995")
```

FIND THE MEDIAN

```
ggplot(medianPackPC,
aes(x=year, y=yearMedian)) +
  geom_point(color="red") +
  geom_line() +
  scale_x_continuous(breaks
= seq(1985,1995)) +
  xlab("Year") + ylab("Median
number of cigarette packs
per capita") +
  ggtitle("Cigarette Packs
Sold in the United States,
from 1985 to 1995")
```



INSIGHTS FROM THE DATA

- What does the data tell you about cigarette usage in 1985 to 1995?
  - The median number of cigarette packs per capita sold in the United State had a large decline from 1985 to 1995.
  - The sharpest decline was from 1989 to 1990.
    - A possible factor: in January, 1989, the 101st Congress passed the Smoking Cost Recovery and Education Tax Act of 1989 that amended the Internal Revenue Code to increase all Federal excise taxes on tobacco products.
    - This bill imposed a new tax of $1.17 per pound on cigarette tobacco manufactured in or imported into the United States.

Source: https://www.congress.gov/bill/101st-congress/house-bill/718?s=1&r=10

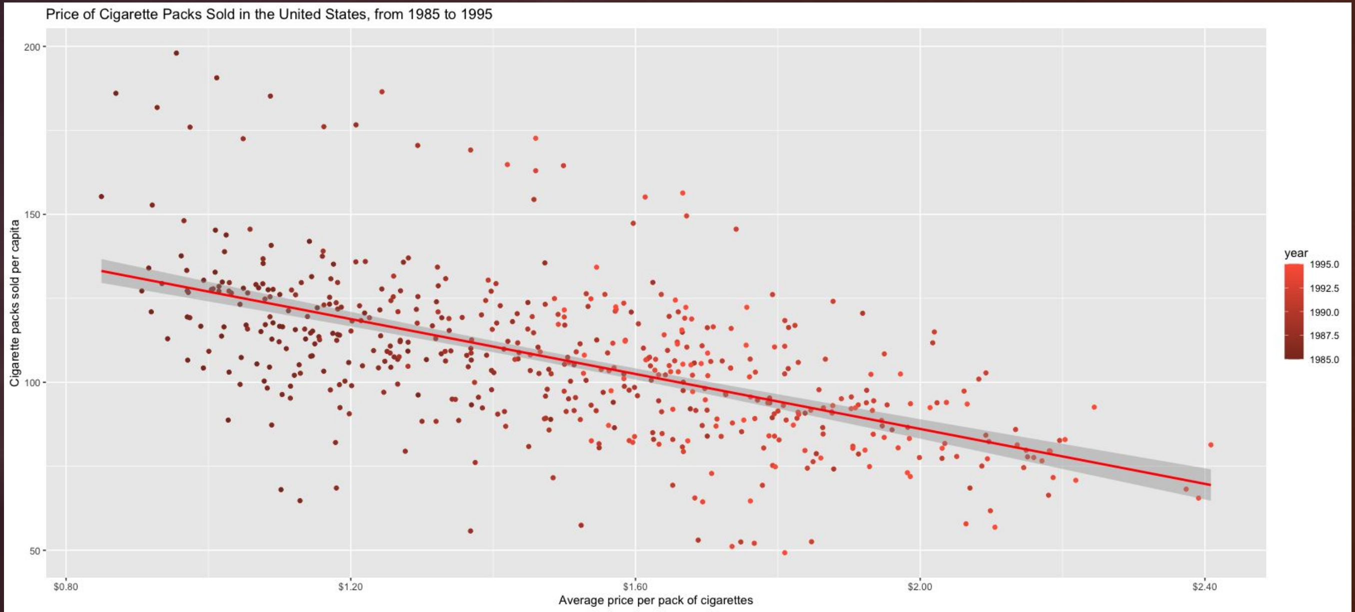
03

CREATE A SCATTER PLOT

Create a scatter plot of **price per pack** vs. **number of packs per capita** for all states and all years.

CREATE A SCATTER PLOT

```
ggplot(Cigarette,
aes(x=avgprs.dollars,
y=packpc, color=year)) +
  geom_point() +

geom_smooth(method=lm,
color="red", se=TRUE) +
  scale_color_gradient(low
= "tomato4", high =
"tomato") +

scale_x_continuous(labels
=scales::dollar_format()) +
  xlab("Average price per
pack of cigarettes") +
ylab("Cigarette packs sold
per capita") +
  ggtitle("Price of Cigarette
Packs Sold in the United
States, from 1985 to 1995")
```



INSIGHTS FROM THE DATA

- How are the price and the per capita packs correlated? Explain why your answer would be expected.

Negatively correlated – as the average price per pack increases, the number of packs per capita decreases.

This makes sense as a correlation, because it follows that a decline in affordability of cigarettes would lead to a decline in consumption.

Until the 1980s, the affordability of cigarettes increased because of the declining real price of cigarettes. Between 1985 and 1990, tobacco manufacturers increased cigarette prices in excess of the rate of inflation and consumer income. Thus, there was a sharp decline in the affordability of cigarettes, although prices remained more affordable than in 1955. The lower affordability of cigarettes in the 1980s corresponds with a decline in consumption.
Source:
https://www.ncbi.nlm.nih.gov/books/NBK236771/

Results: Tobacco company documents provide clear evidence on the impact of cigarette prices on cigarette smoking, describing how tax related and other price increases lead to significant reductions in smoking, particularly among young persons.
Source:
https://tobaccocontrol.bmj.com/content/11/suppl_1/i62

In more recent years,
- data suggests that increasing the price of tobacco is the single most effective way to reduce consumption.
- a 10% increase in price of tobacco is estimated to reduce overall cigarette consumption by 3-5%.

Source:
https://www.cdc.gov/tobacco/data_statistics/fact_sheets/economics/econ_facts/index.htm

INSIGHTS FROM THE DATA

- Does the relationship between the average price per pack vs. packs per capita change over time?

The relationship between **average price per pack** and **packs per capita** seems to keep a negative correlation over the measured years, but it is only a **moderate correlation**, confirmed by **cor.test( )** with a **correlation of approximately -0.585**.

~~~
```r
cor.test(Cigarette$avgprs,
Cigarette$packpc,
method="pearson", use =
"complete.obs")
```
~~~

~~~
```text
Pearson's product-moment correlation

data:  Cigarette$avgprs and Cigarette$packpc
t = -16.562, df = 526, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6388606 -0.5264104
sample estimates:
      cor
-0.5854443
```
~~~

INSIGHTS FROM THE DATA

○ Do a linear regression for these two variables. How much variability does the line explain?

Variability: about 34%

- Adjusted R-squared of the line: 0.3415
- average price per pack is able to explain about 34% of the factors that go into number of packs sold per capita. The rest is covered by other variables that have not been included in the model.

04

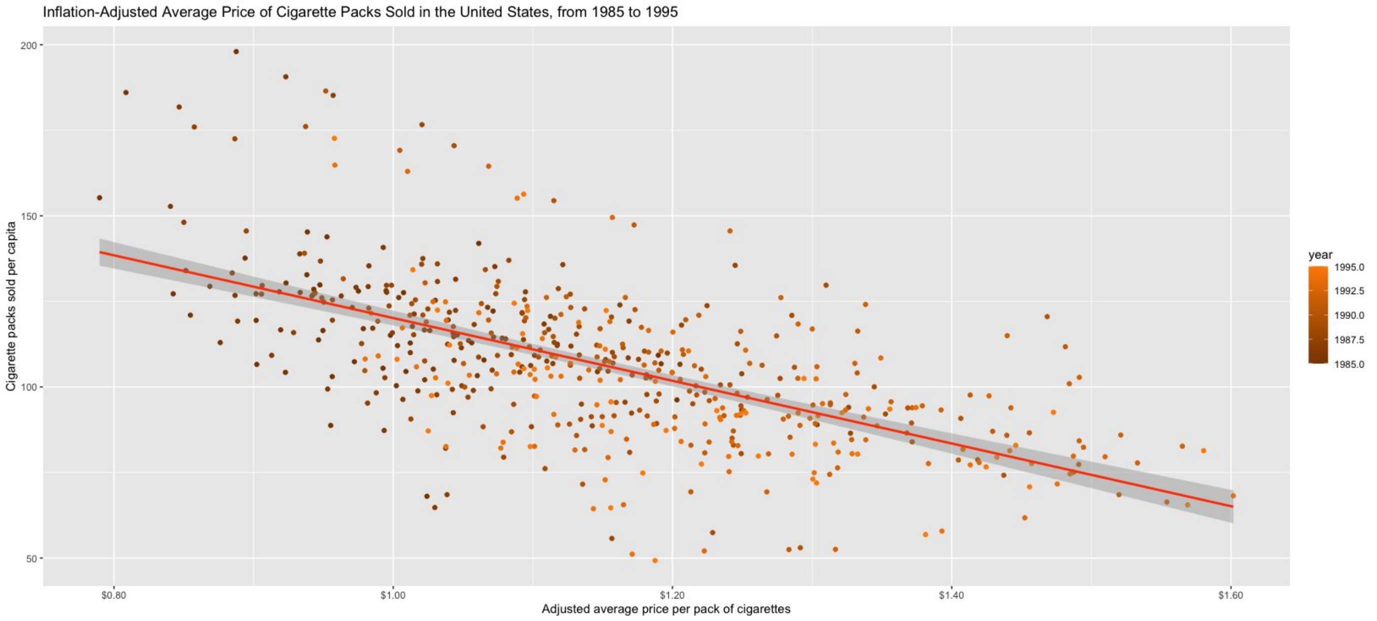
ADJUST FOR INFLATION

Adjust for inflation by dividing the **average price per pack** by **cpi** (consumer price index for the year)

CREATE A SCATTER PLOT

```
Cigarette <-
mutate(Cigarette,
inflation.dollars =
(avgprs/cpi)/100)
```

```
# Plot with this adjusted price
ggplot(Cigarette,
aes(x=inflation.dollars,
y=packpc, color=year)) +
  geom_point() +
  geom_smooth(method=lm,
color="orangered1", se=TRUE) +
  scale_color_gradient(low =
"darkorange4", high =
"darkorange") +

scale_x_continuous(labels=scal
es::dollar_format()) +
  xlab("Adjusted average price
per pack of cigarettes") +
ylab("Cigarette packs sold per
capita") +
  ggtitle("Inflation-Adjusted
Average Price of Cigarette
Packs Sold in the United
States, from 1985 to 1995")
```



INSIGHTS FROM THE DATA

- What is the difference in variability once price is adjusted for inflation?

  Linear regression with adjusted price

  Variability: about 37% (vs. unadjusted: 34%)

  - Adjusted R-squared of the line: 0.3757

05

COMPARE 1985 to 1995 WITH DEPENDENT T-TEST

Use a dependent T-test to compare the number of cigarette packs sold by capita in the year 1985 and the year 1995.

PREPARING THE DATA

# Create a data frame with just the rows from 1985.
```r
Cigarette.1985 <-
filter(Cigarette, year == 1985)
```

# Create a data frame with just the rows from 1995.
```r
Cigarette.1995 <-
filter(Cigarette, year == 1995)
```

Create vector of number of packs per capita from 1985
```r
packspc.1985 <-
Cigarette.1985$packpc
```

Create vector of number of packs per capita from 1995
```r
packspc.1995 <-
Cigarette.1995$packpc
```

PAIRED T-TEST

Use a paired t-test to see if the number of packs per capita in 1995 was significantly different than the number of packs per capita in 1985.

Hypothesis Test

Population 1 is the number of packs per capita in 1985
Population 2 is the number of packs per capita in 1995

$H_0 : \mu_2 - \mu_1 = 0$
$H_a : \mu_2 - \mu_1 \neq 0$
- the samples are paired

```r
t_dep <- t.test(packspc.1985, packspc.1995, paired = TRUE)
t_dep
```

```text
Paired t-test
data:  packspc.1985 and packspc.1995
t = 14.789, df = 47, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 22.21151 29.20576
sample estimates:
mean difference
       25.70863
```

INSIGHTS FROM THE DATA

- Is there a significant difference in the number of packs per capita in 1985 vs. 1995
  - p-value < 0.05 (default alpha=0.05)
  - The data presents strong evidence to REJECT the null hypothesis.
  - The data indicates a significant difference
  - The data presents strong evidence that the difference between the number cigarette packs sold per capita in 1995 is significantly different than the number of cigarette packs sold per capita in 1985.

05+

GRAPHING DATA OF THE DEPENDENT T-TEST

Use a dependent t-test to compare the number of cigarette packs sold by capita in the year 1985 and the year 1995.

PREPARING THE DATA

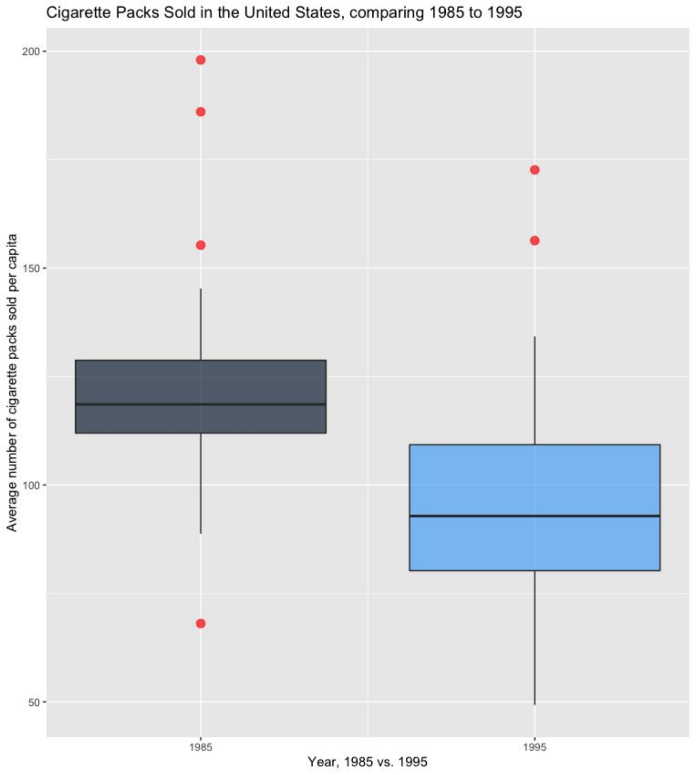
~~~
# Create data frame with
select of only year, and packpc
with filter of both 1985 and
1995 at the same time.

```r
Cigarette.1985v1995 <-
Cigarette %>% select(year,
packpc) %>% filter(year %in%
c(1985, 1995))
View(Cigarette.1985v1995)
```
~~~

~~~
# Create box plot – simple.
```r
ggplot(Cigarette.1985v1995, aes(x = year, y = packpc)) +
geom_boxplot(aes(group=year))
  ```
~~~

(work continued on next page...)

GRAPHING THE DATA

~~~
 # Create box plot – prettier.
```r
ggplot(Cigarette.1985v1995, aes(x = year, y = packpc,
group = year, fill = year)) +
  geom_boxplot(
    # additional color settings
    alpha=0.7,
    # custom outliers
    outlier.color="red",
    outlier.fill="red",
    outlier.size=3
      ) +
  scale_x_continuous(breaks = c(1985,1995)) +
  theme(legend.position="none") +
  xlab("Year, 1985 vs. 1995") + ylab("Average number of
cigarette packs sold per capita") +
  ggtitle("Cigarette Packs Sold in the United States,
comparing 1985 to 1995")
```
~~~


Cigarette Packs Sold in the United States, comparing 1985 to 1995

06

INDEPENDENT ANALYSIS:

What is the relationship between average income and cigarette pack sales per capita?

PREPARING THE DATA

○ What variables do we need?

- **income** - total state personal income
- **pop** - state population
- **packpc**

○ How to calculate income per capita?

Use mutate() to add a column with calculation.

From Bureau of Economic Analysis, seems that total state income is usually presented in millions (Source: https://www.bea.gov/news/2022/personal-income-state-1st-quarter-2022)

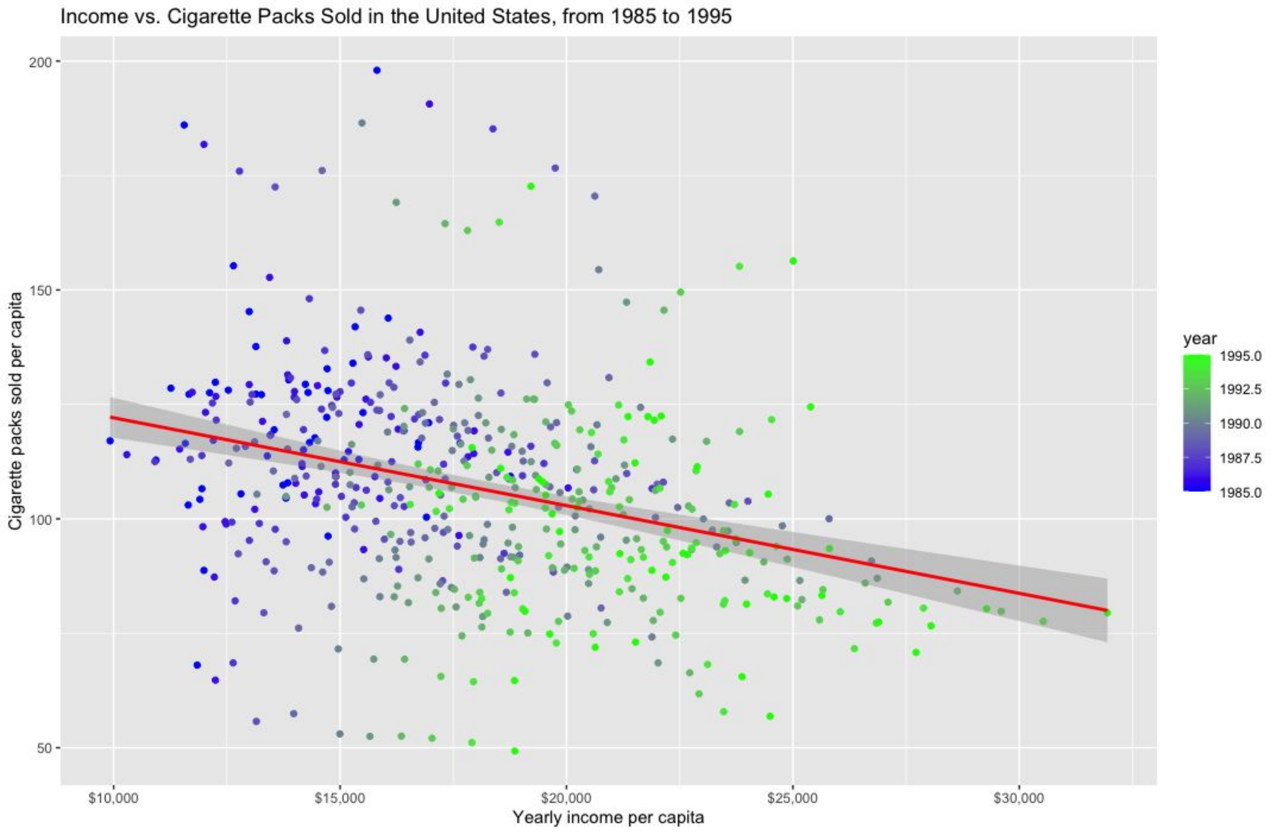
Calculate with: `(income * 1000)/pop`

Confirmed by checking Alabama per capita personal income

(Source: https://fred.stlouisfed.org/series/ALPCPI)

CREATE A SCATTER PLOT

~~~
# Plot with best fit line

```r
ggplot(Cigarette,
aes(x=incomepercapita,
y=packpc, color=year)) +
  geom_point() +
  geom_smooth(method=lm,
color="red", se=TRUE) +
  scale_color_gradient(low =
"blue", high = "green") +

scale_x_continuous(labels=scal
es::dollar_format()) +
  xlab("Yearly income per
capita") + ylab("Cigarette packs
sold per capita") +
  ggtitle("Income vs. Cigarette
Packs Sold in the United
States, from 1985 to 1995")
```
~~~



INSIGHTS FROM THE DATA

- Linear Regression to find variability
  - Variability: 10%
  - Adjusted R-squared of the line: 0.1009
  - The **income per capita** does not explain variability as much as the other factor we looked at, **average price per pack**.
  - **average price per pack** is able to explain about 34% (nominal) / 37% (adjusted for inflation) of the factors that go into **number of packs sold per capita**, vs. only **income per capita** at 10%.

07

INDEPENDENT ANALYSIS:

What was the packs sold per capita in state of TN vs. the rest of the United States?

WHY THIS QUESTION?

- The state of TN has a long history of tobacco farming
- This time of year is tobacco harvesting time — along with “smoking the tobacco” in barns, filling nearby areas with the smell of tobacco smoke
- My husband's paternal grandfather died from lung cancer after a lifelong habit of smoking, and it was partially what led my husband to refuse to pursue agriculture as a profession.
- I was interested in seeing the sale of cigarettes in TN vs. the United States as a whole, from 1985 to 1995.

PREPARING THE DATA

○ What variables do we need?

- **packpc**
- new: **packpc.TN** <- By year, the **packpc** for just TN
- new: **packpc.notTN** <- By year, the **packpc** for all states excluding TN

PREPARING THE DATA

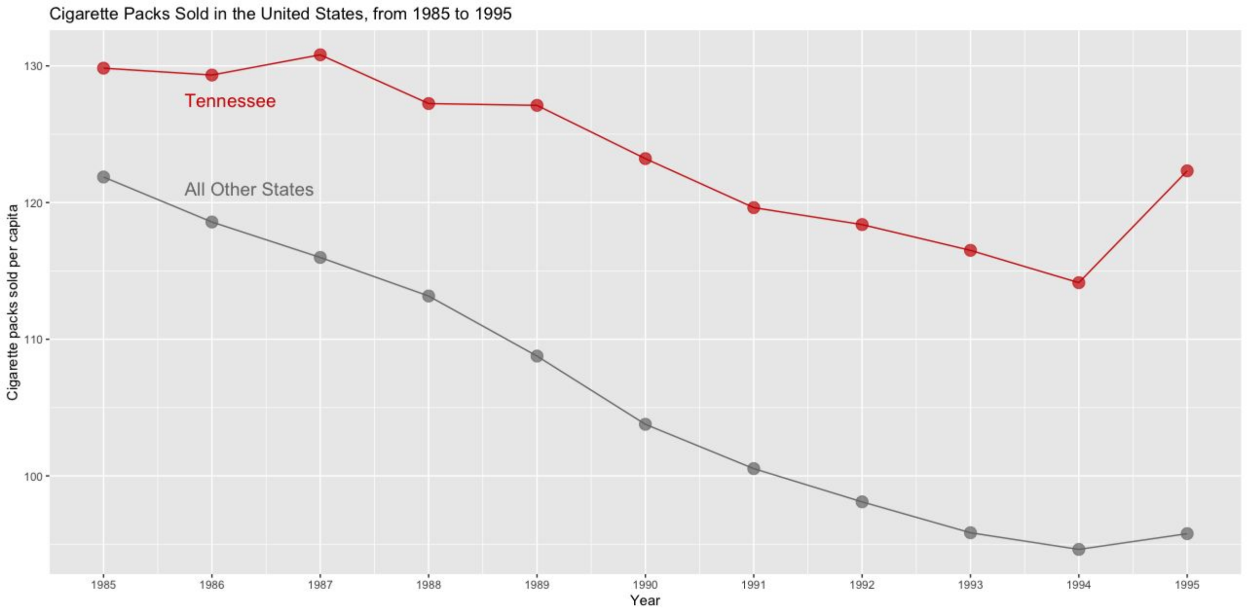
```
# Create data frame with only TN

packpc.TN <- Cigarette %>% select(state, year, packpc) %>%
filter(state=="TN")
View(packpc.TN)
# remove "state" column
packpc.TN <- packpc.TN %>% select(year, packpc)

# get column names from packpc.TN and rename
colnames(packpc.TN)
colnames(packpc.TN) <- c("year", "TN_Mean")
```

```
# Create data frame with all state excluding
TN
# and average (mean) the packpc for all states
by year

packpc.notTN <- Cigarette %>% select(state,
year, packpc) %>% filter(state != "TN") %>%
group_by(year) %>% summarize(stateMean =
mean(packpc))

View(packpc.notTN)
```

```
# Using rbind() to concatenate the 2 data frames
packpc.byState <- list(packpc.TN, packpc.notTN)
packpc.byState %>% reduce(full_join, by="year")
packpc.byState <- as.data.frame(packpc.byState)
```

CREATE A SCATTER PLOT

```
# Plot together
p <- ggplot(packpc.byState,
aes(x=year))+
  geom_line(aes(y=stateMean),
color="grey46") +
  geom_point(aes(y=stateMean),
color="grey46", size=4, alpha=0.7)
+
  geom_line(aes(y=TN_Mean),
color="red3") +
  geom_point(aes(y=TN_Mean),
color="red3", size=4, alpha=0.7) +
  theme(legend.position="none")
+
  scale_x_continuous(breaks =
seq(1985,1995)) +
  xlab("Year") + ylab("Cigarette
packs sold per capita") +
  ggtitle("Cigarette Packs Sold in
the United States, from 1985 to
1995")

# Add the labels
p + annotate("text",
x=c(1985.75,1985.75), y=c(127.5,121),
label=c("Tennessee", "All Other
States"), color=c("red3", "grey46"),
size=5, hjust=0)
```



INSIGHTS FROM THE DATA

- How does TN compare to the rest of the United States for this measurement?
    - TN has a higher number of cigarettes sold per capita every year measured.
    - TN did not have as steep a decline as the combined rest of the United States.
    - TN had a slight spike in 1987, where the rest of the states did not.
    - TN had a much steeper spike in 1995 than the rest of the states.

THANKS!