Prevalence of Cigarette Use in the United States

Alberta “Albi” Kovatcheva

Background

- The ‘Cigarettes’ data set in R:
  - A panel of 48 observations from 1985 to 1995
  - Number of observations: 528
  - Observation: regional
  - Country: United States (Continental)

Scenario

- A researcher wants to explore the prevalence of cigarette use in the United States.

Tasks

Create a boxplot of the average number of packs per capita by state. Which states have the highest number of packs? Which have the lowest?

Find the median over all the states of the number of packs per capita for each year. Plot this median value for the years from 1985 to 1995. What can you say about cigarette usage in these years?

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

Are the price and the per capita packs positively correlated, negatively correlated, or uncorrelated? Explain why your answer would be expected.

Change your scatter plot to show the points for each year in a different color. Does the relationship between the two variable change over time?

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

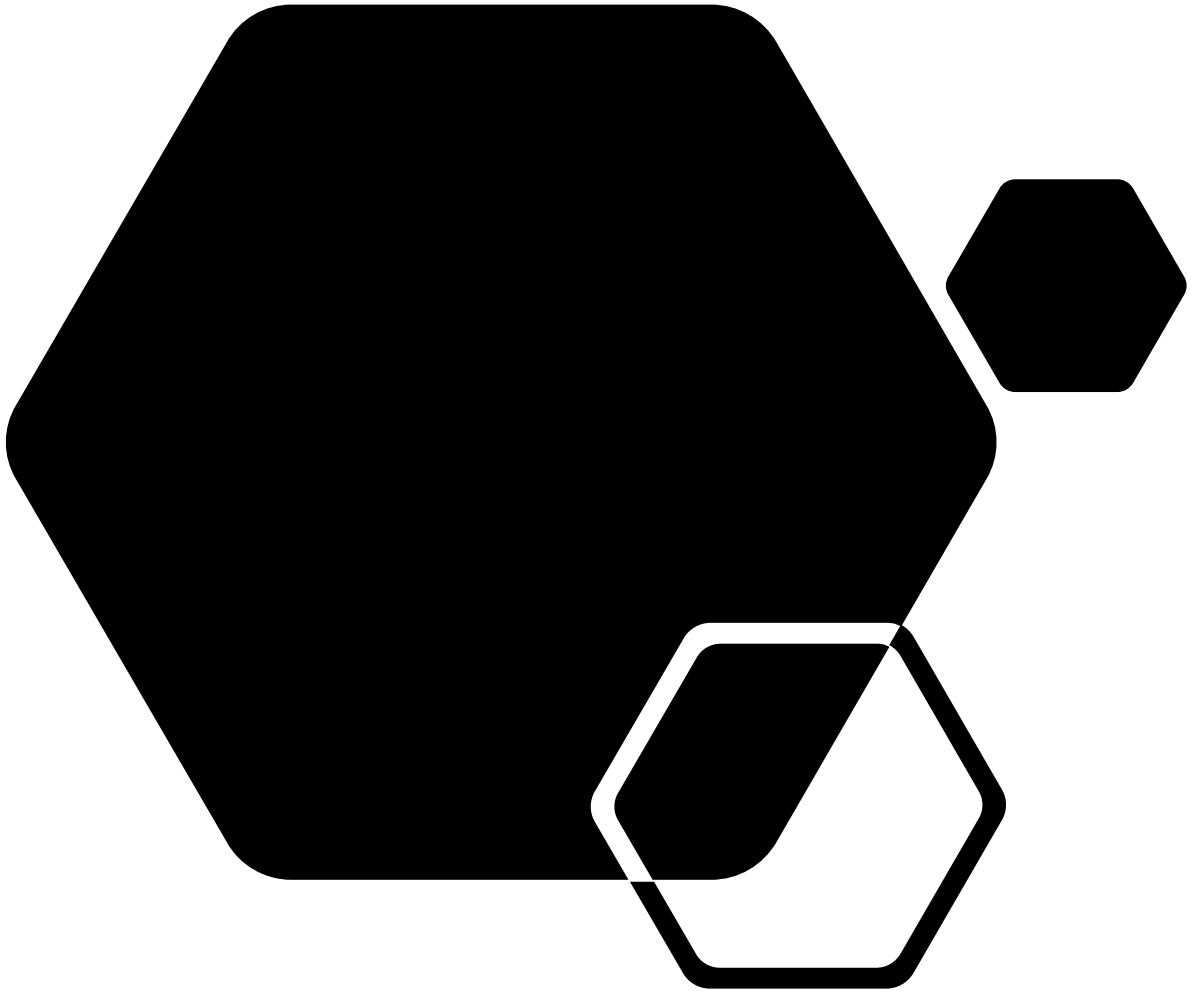
The plot above does not adjust for inflation. You can adjust the price of a pack of cigarettes for inflation by dividing the avgprs variable by the cpi variable. Create an adjusted price for each row, then re-do your scatter plot and linear regression using this adjusted price.

Create a data frame with just the rows from 1985. Create a second data frame with just the rows from 1995. Then, from each of these data frames, get a vector of the number of packs per capita. 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.

In the process of doing this project, have any questions come to mind that this data set could answer? If so, pick one and do the analysis to find the answer to your question.

Part 1

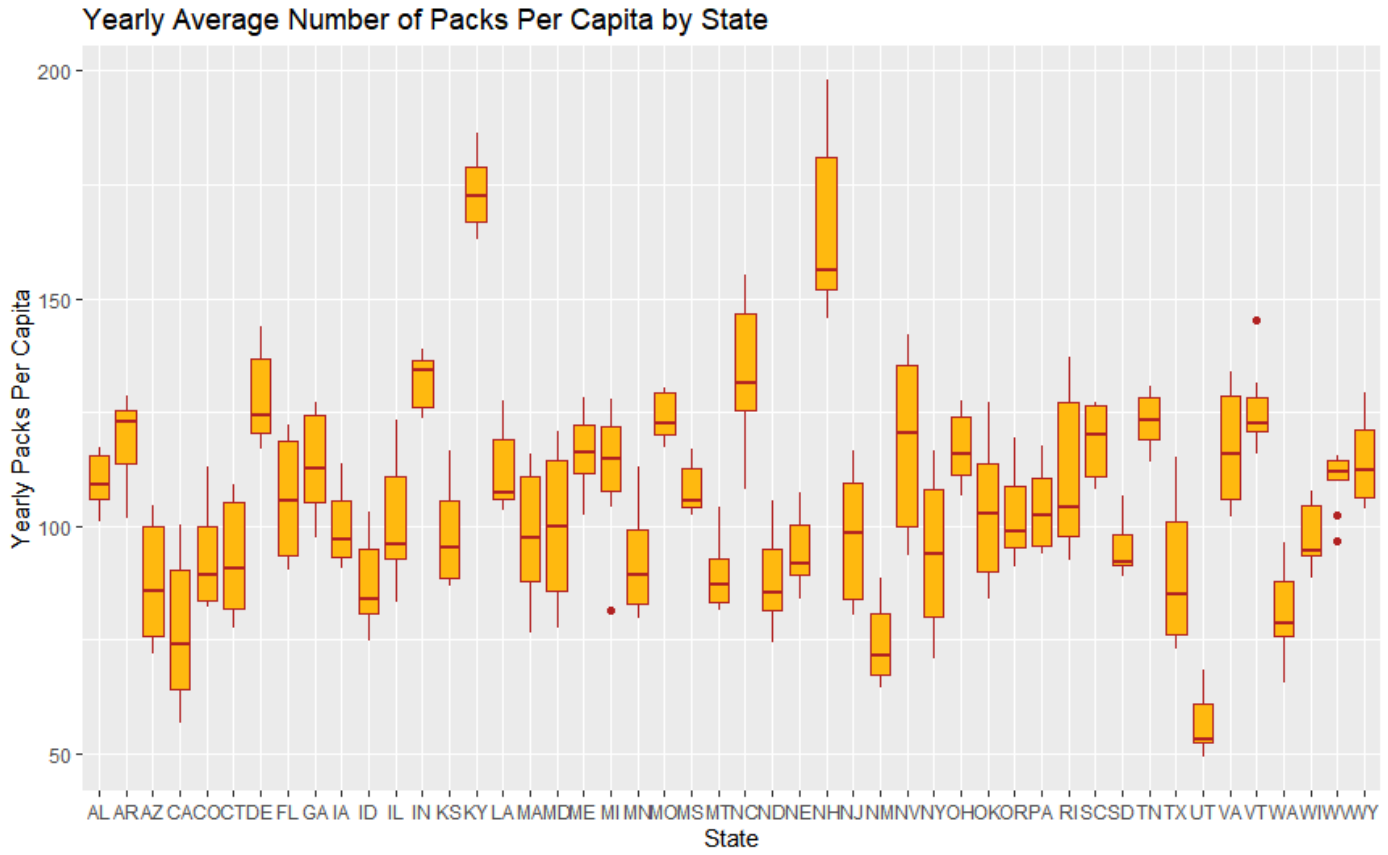
Yearly Number of Packs Per Capita by State



Yearly Number of Packs Per Capita by State

Code to create this boxplot:

```
ggplot(Cigarette, aes(x = state, y = packpc)) +
  geom_boxplot(fill = "darkgoldenrod1", color = "firebrick") +
  xlab("State") + ylab("Yearly Packs Per Capita") +
  ggtitle("Yearly Average Number of Packs Per Capita by State")
```



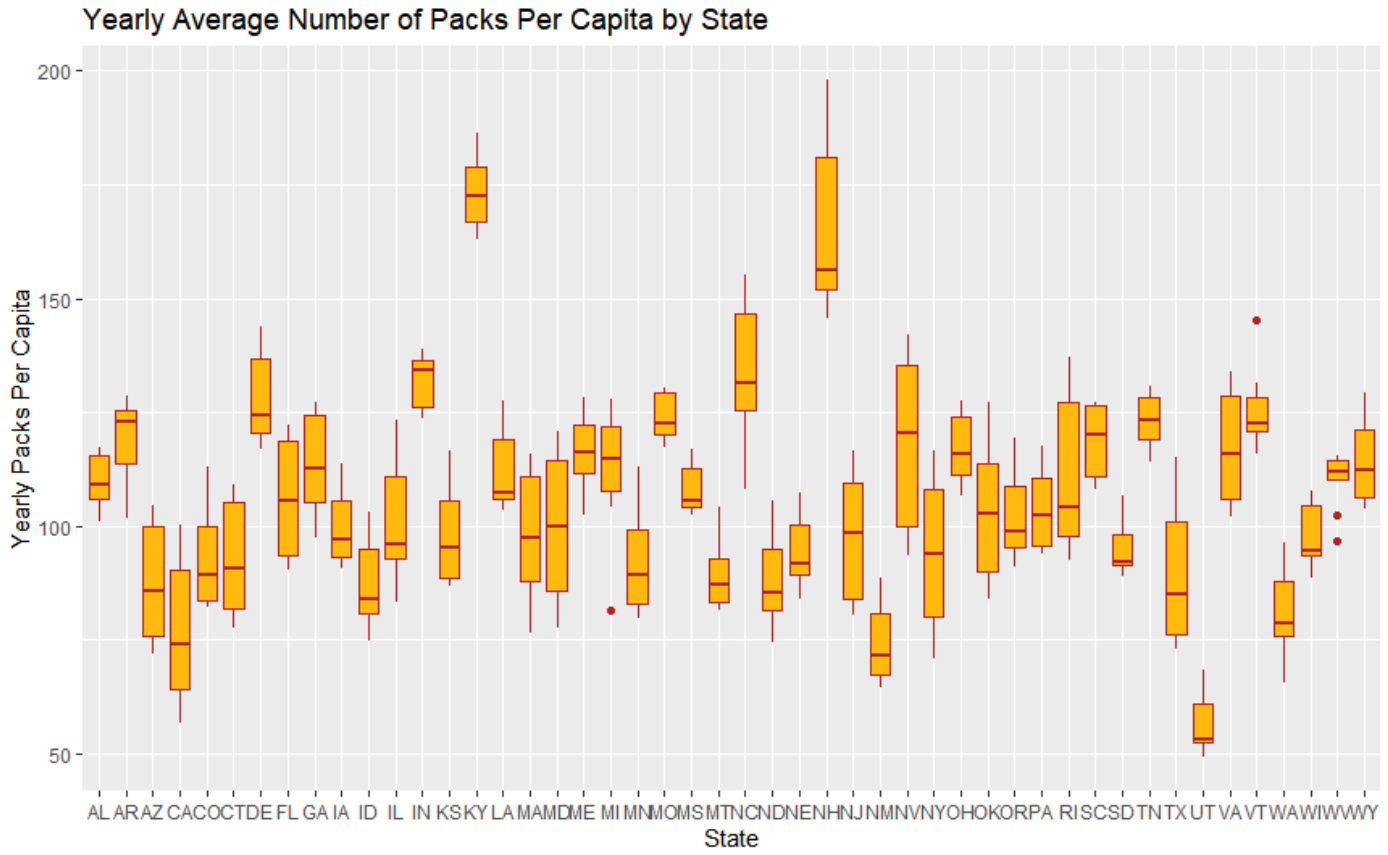
Analysis

- Out of all 50 states in the U.S.A., Kentucky has the highest average number of packs per capita (173.90494), and Utah has the lowest (56.82223).

Code to gather this information:

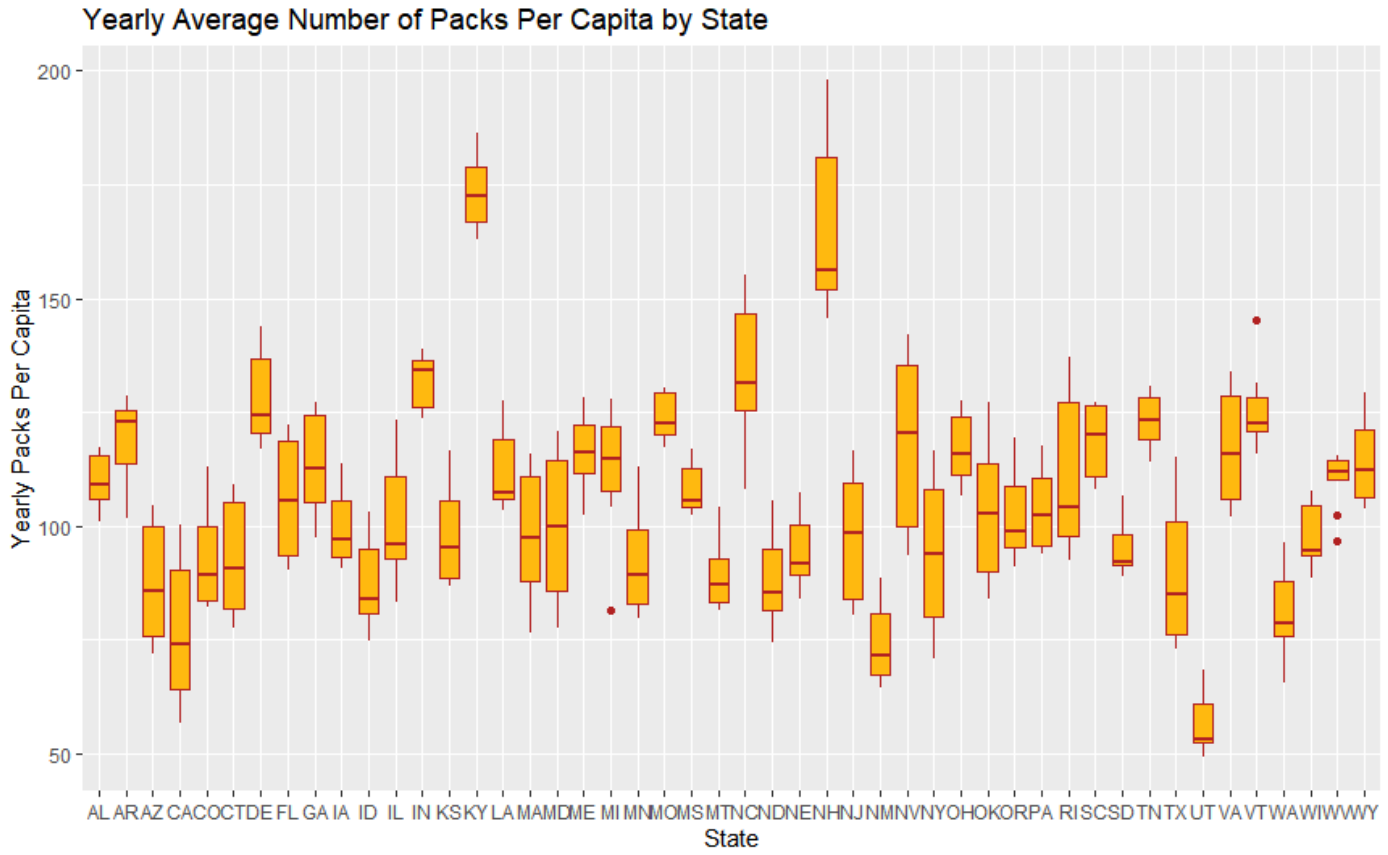
```
Cigbystate <- Cigarette %>%
group_by(state) %>% summarize(mean=
mean(packpc)) %>% arrange(desc(mean))

View(Cigbystate)
```



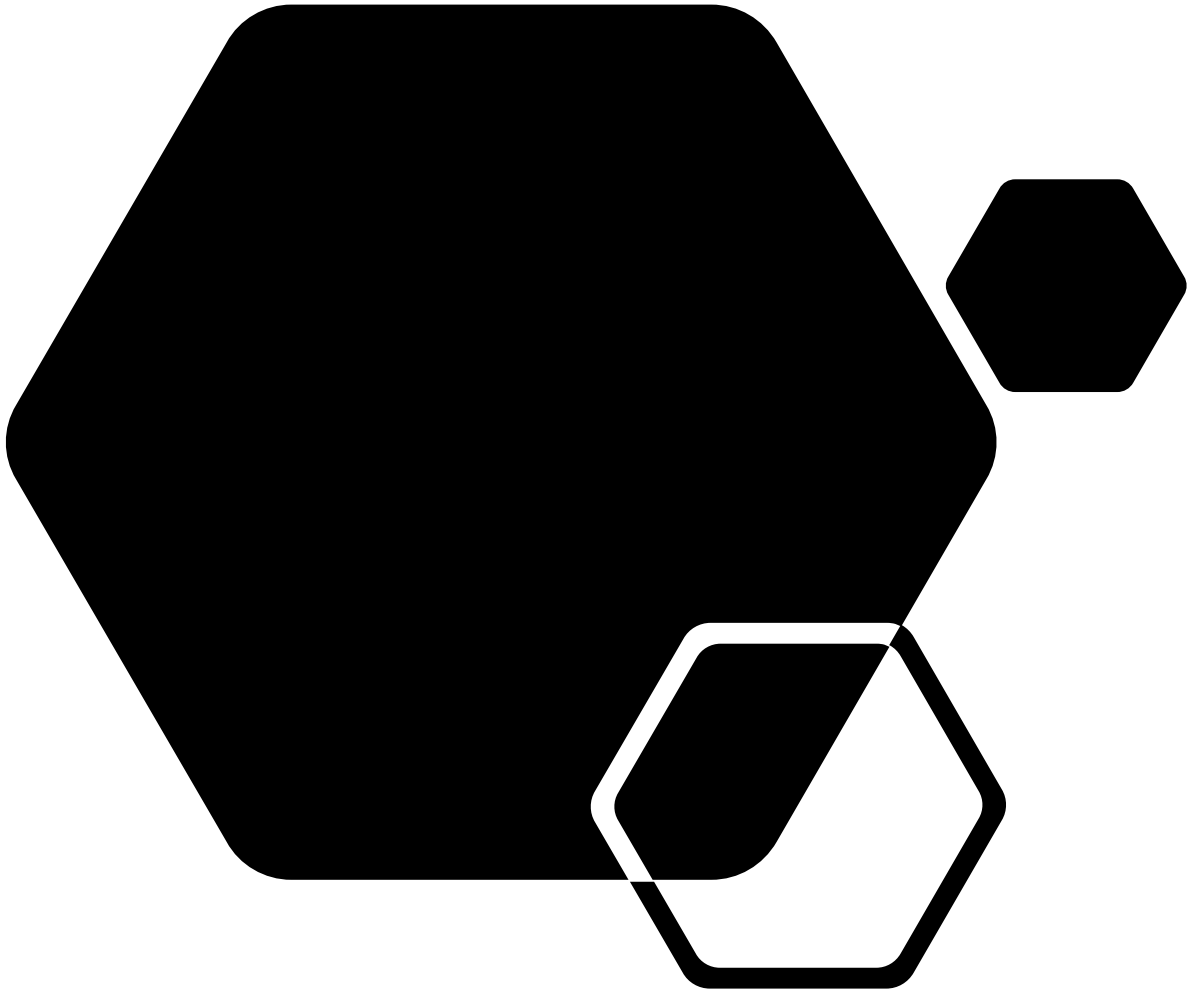
Analysis

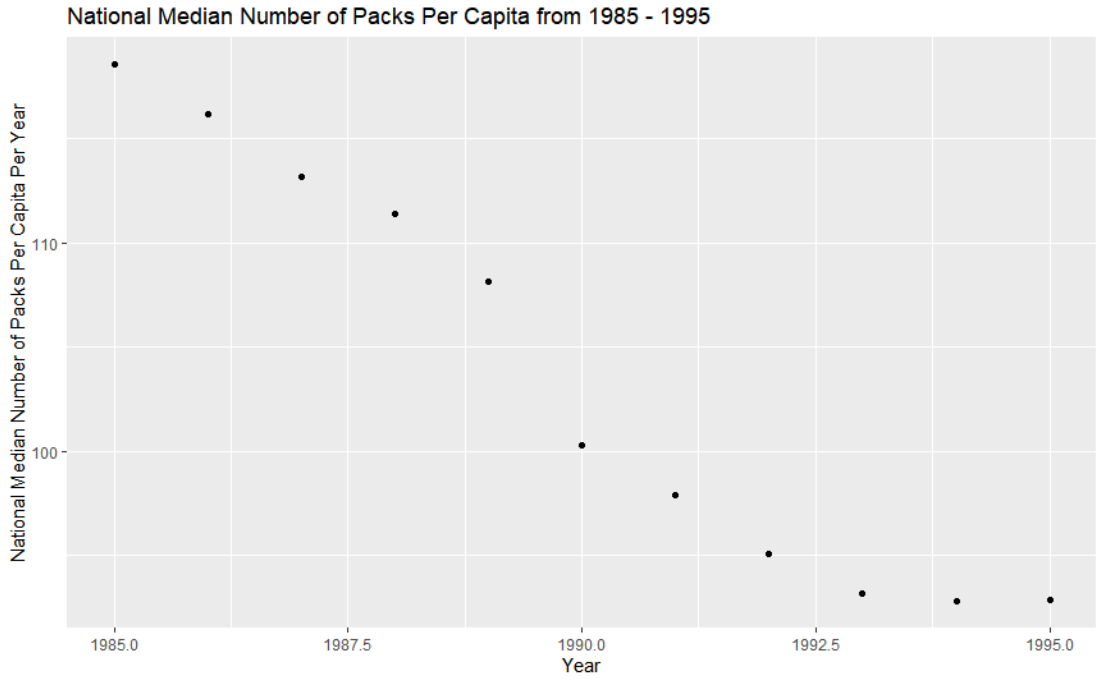
- In the U.S., Utah has the lowest average number of packs per capita (56.82223).
- This could be because Utah taxes tobacco products more heavily than most other states.
- According to the Sales Tax Book, in Utah, cigarettes are subject to a state excise tax of $1.70 per pack of 20. Cigarettes are also subject to Utah sales tax of approximately $0.36 per pack, which adds up to a total tax per pack of $2.06. Therefore, many residents buy cigarettes from illegal, out-of-state vendors. These illegal purchases would be reflected in this data.



Part 2

Median Number of Packs Per Capita for Each Year



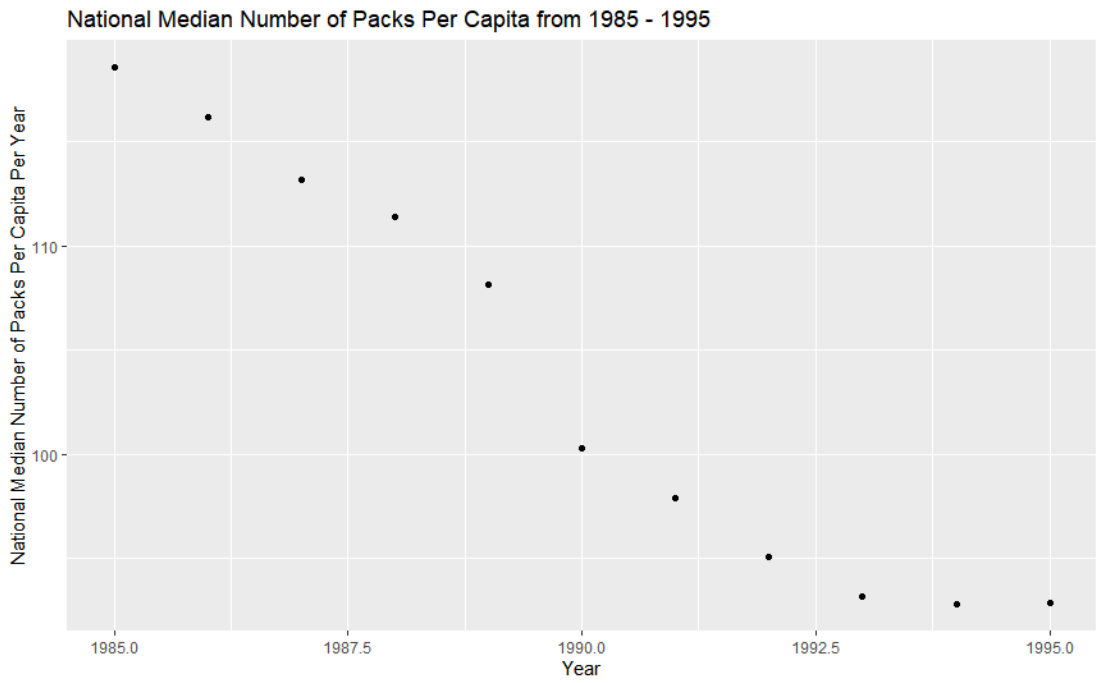


YEAR	MEDIAN
1985	118.5784
1986	116.2005
1987	113.1733
1988	111.3792
1989	108.1627
1990	100.3012
1991	97.8923
1992	95.07264
1993	93.18804
1994	92.79489
1995	92.83718

Median Number of Packs Per Capita for Each Year

Code to create this scatter plot:

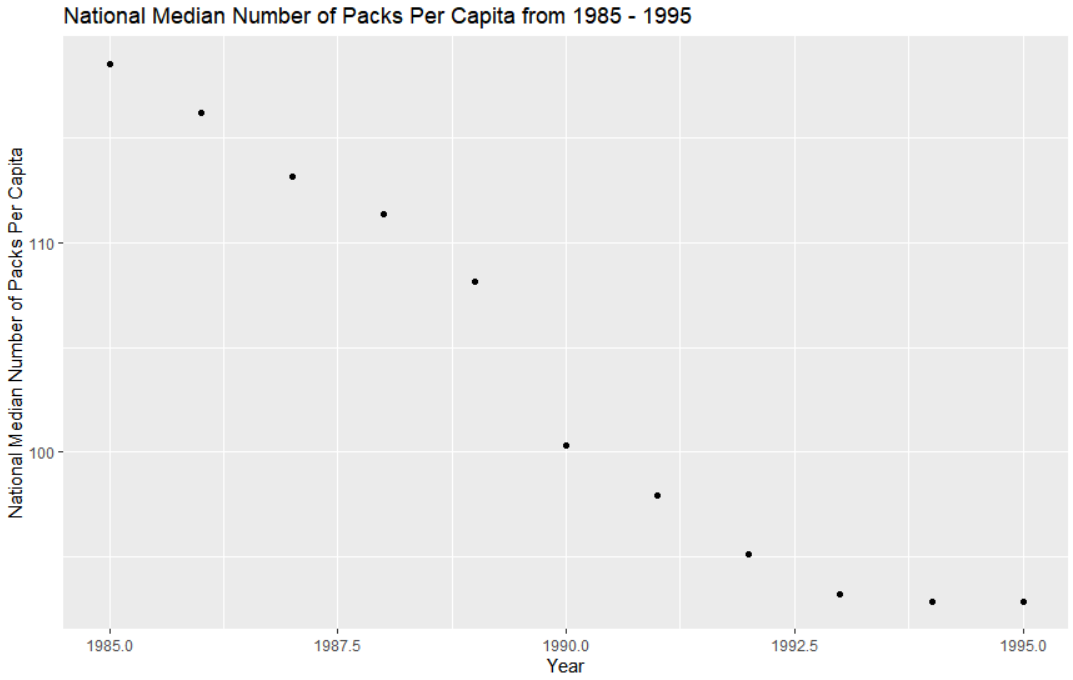
```
unique(Cigarette$year)
 ggplot(Cigmedian, aes(x = year, y = median)) + geom_point() +
 ggtitle("National Median Number of Packs Per Capita from 1985 - 1995") +
 xlab("Year") +
 ylab("National Median Number of Packs Per Capita Per Year")
```



YEAR	MEDIAN
1985	118.5784
1986	116.2005
1987	113.1733
1988	111.3792
1989	108.1627
1990	100.3012
1991	97.8923
1992	95.07264
1993	93.18804
1994	92.79489
1995	92.83718

Analysis

- From 1985 to 1995, national cigarette use has been consistently declining.
- The decline of national cigarette use has been undergoing an approximately logistic decay, with the most significant decrease occurring between 1989 and 1990.



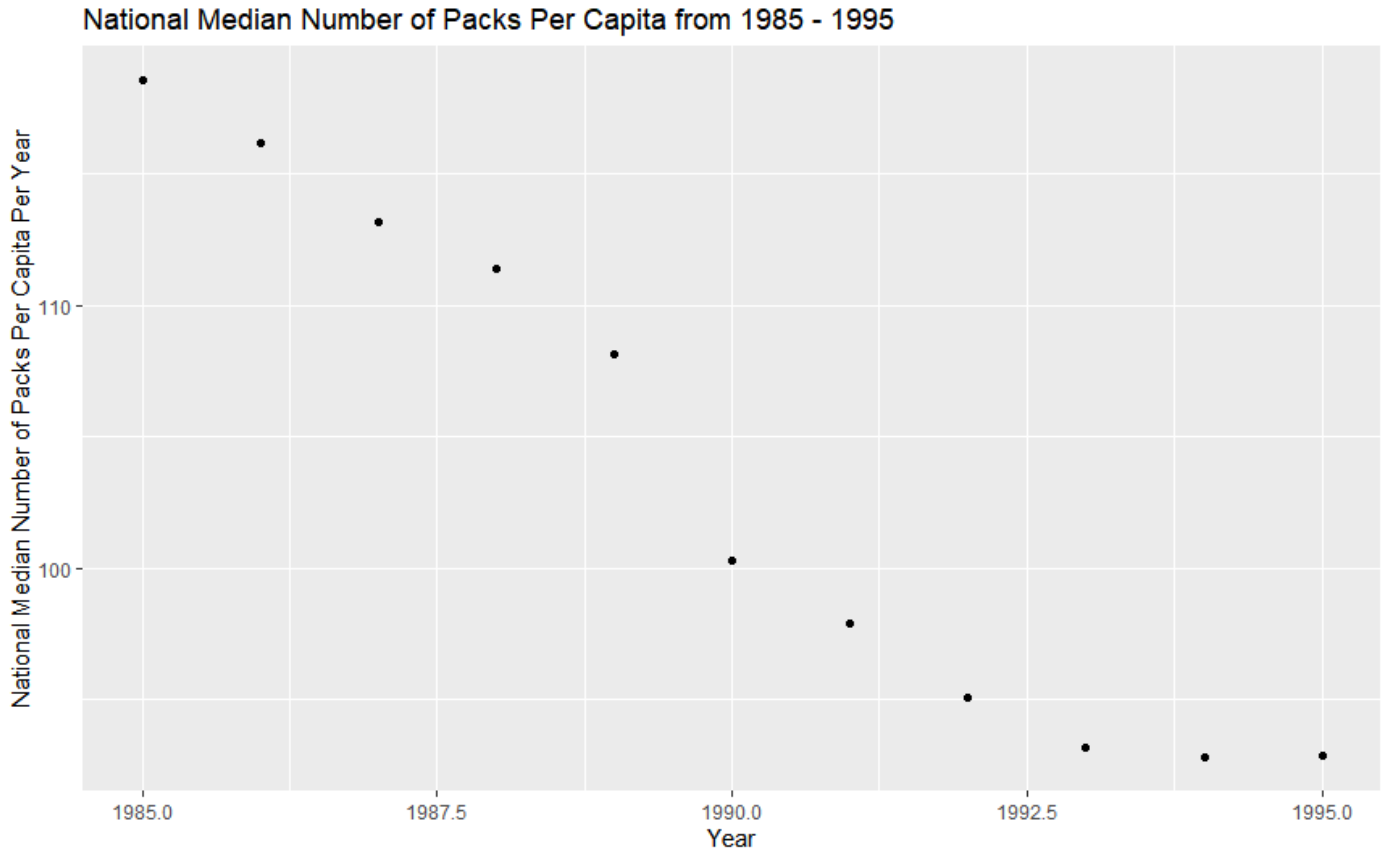
YEAR	MEDIAN
1985	118.5784
1986	116.2005
1987	113.1733
1988	111.3792
1989	108.1627
1990	100.3012
1991	97.8923
1992	95.07264
1993	93.18804
1994	92.79489
1995	92.83718

Median Number of Packs Per Capita for Each Year

- From 1985 to 1995, national cigarette use has been consistently declining.
- The decline of national cigarette use has been undergoing an approximately logistic decay with the most significant decrease in occurring between 1989 and 1990. During this period, the median number of packs per capita dropped from approximately 108 to 100.

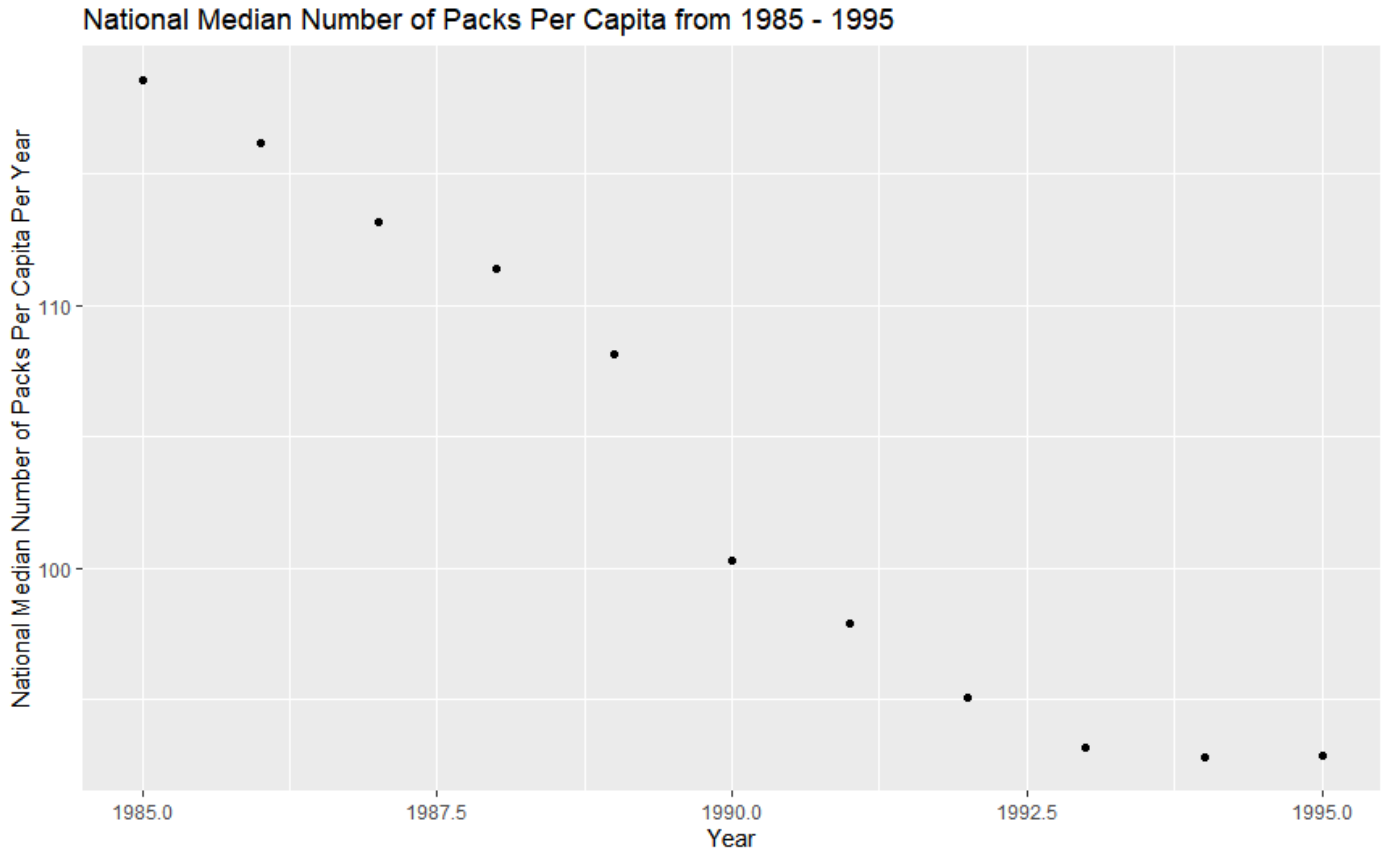
Analysis

- According to Tobacco Control, The California Tobacco Control Program (CTCP) is one of the longest-running comprehensive tobacco control programs in the USA. The CTCP began because of a November 1988 ballot initiative known as Proposition 99, which added a 25-cent tax per cigarette pack and a proportional tax increase to other tobacco products beginning 1 January 1989. Similar taxes were levied by other states to fund public health programs to prevent and reduce tobacco use, provide healthcare services, fund tobacco-related research and protect environmental resources.



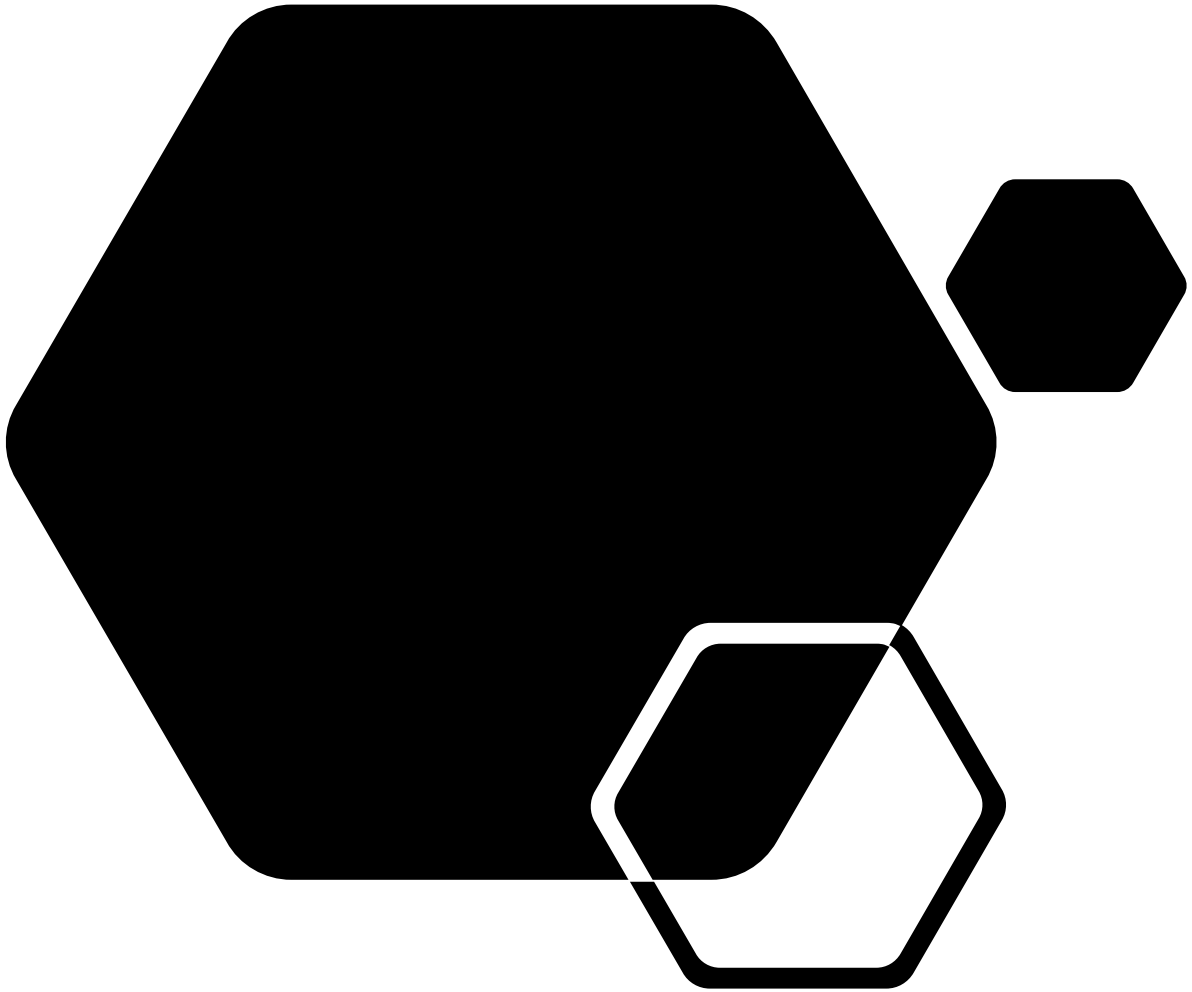
Analysis

- Cigarette usage plateaued from 1993 to 1995.
- According to the CDC, the sharp decline in cigarette use would have continued had it not been for increased market shares of discount cigarette brands and significantly increased national cigarette marketing.



Part 3

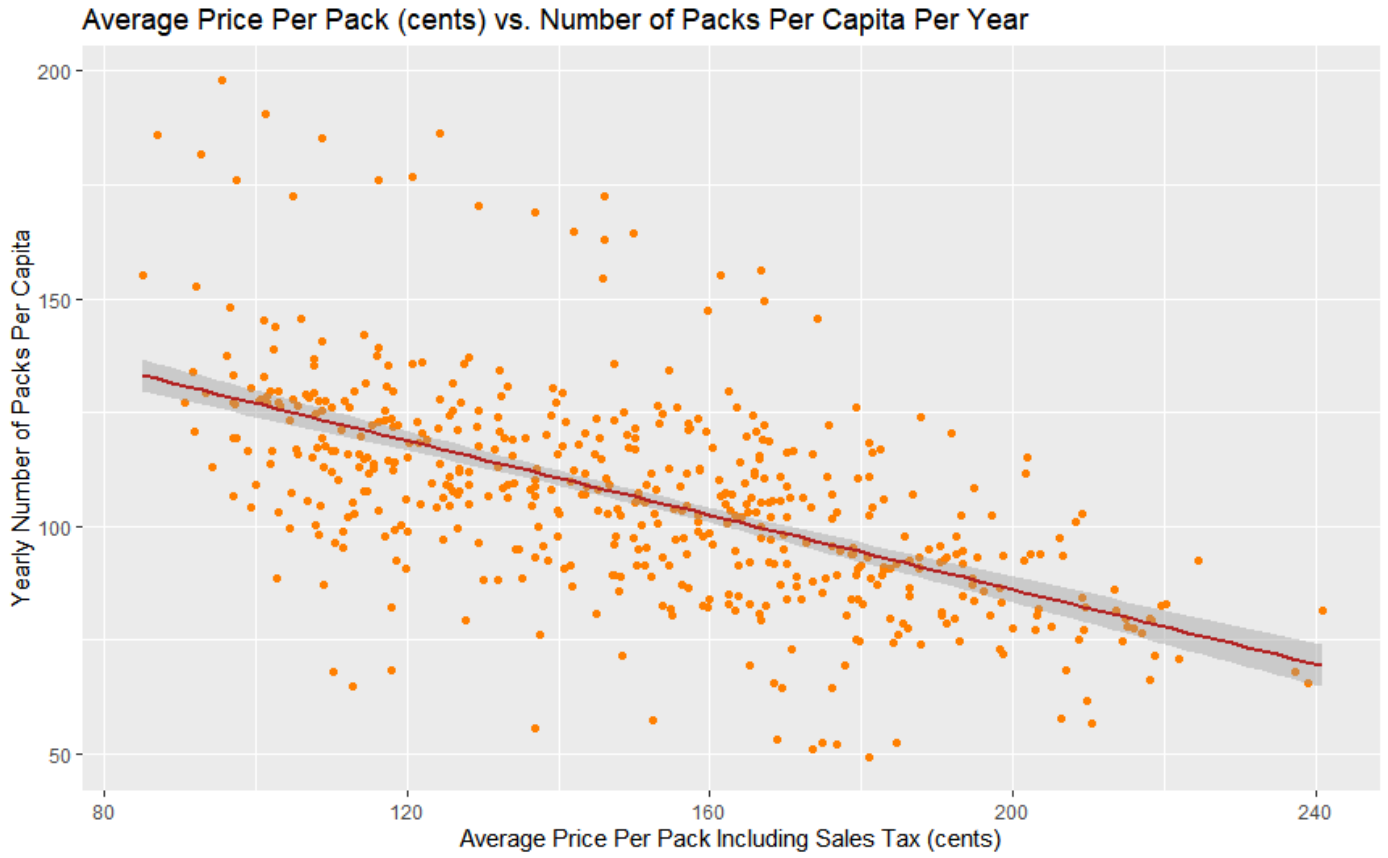
Average Price Per Pack vs. Yearly Number Of Packs Per Capita



Average Price Per Pack vs. Yearly Number Of Packs Per Capita

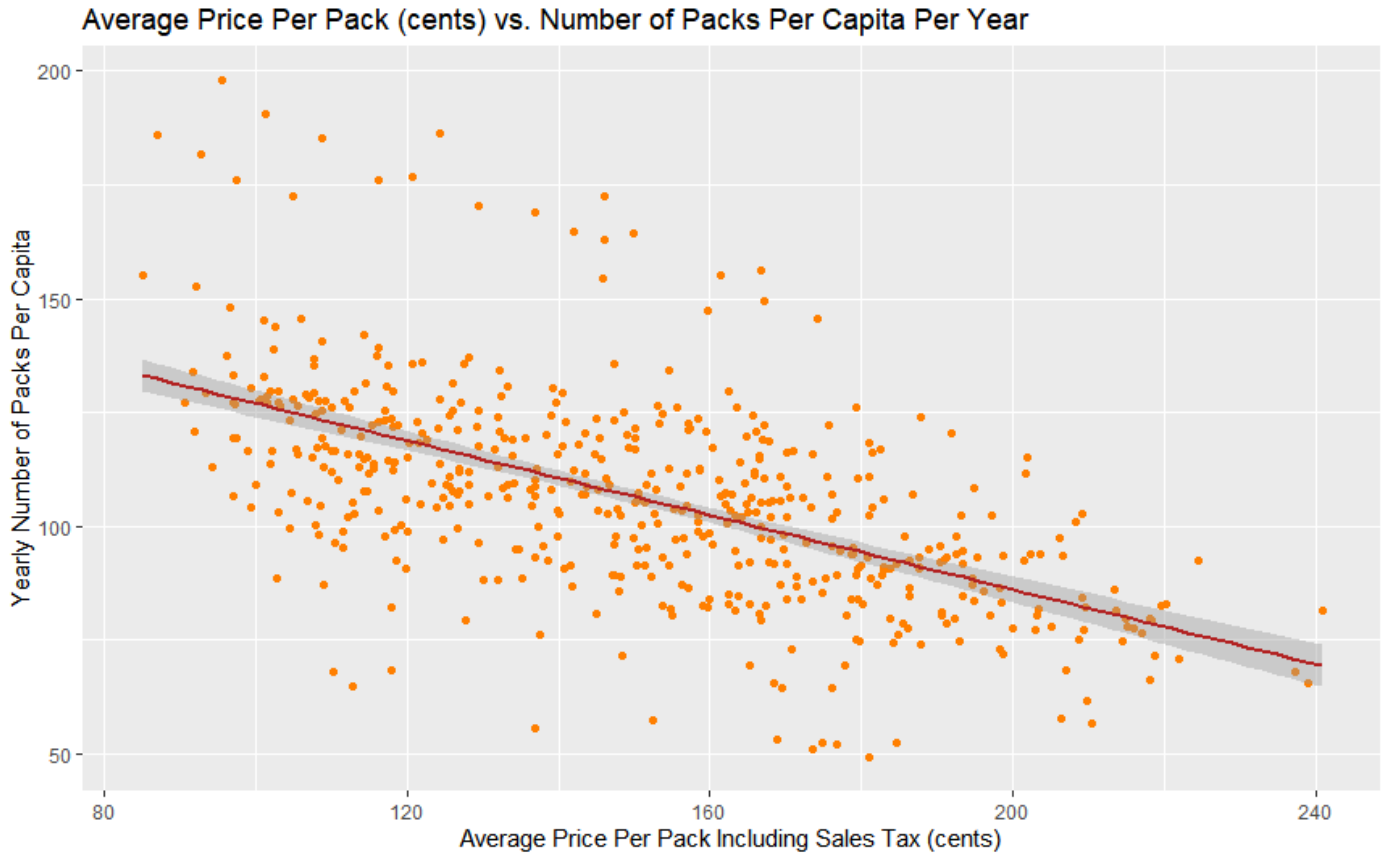
Code to create this scatter plot:

```
ggplot(Cigarette, aes(x = avgprs, y = packpc)) +
geom_point(color = "darkorange1") +
geom_smooth(method = lm, color =
"firebrick") +
  xlab("Average Price Per Pack Including Sales
Tax (cents)") +
  ylab("Yearly Number of Packs Per Capita") +
  ggtitle("Average Price Per Pack (cents) vs.
Number of Packs Per Capita Per Year")
```



Analysis

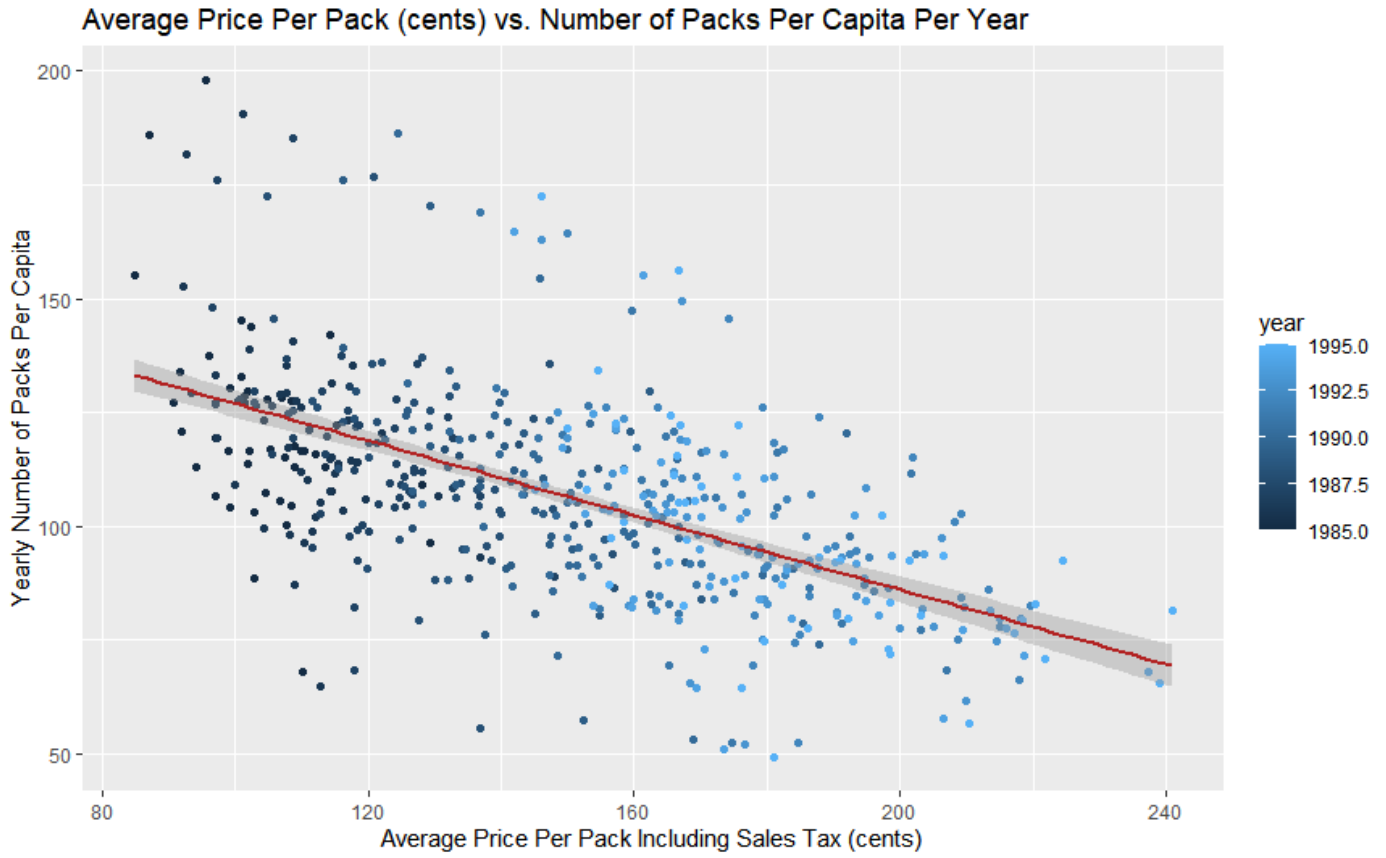
- There appears to be a negative correlation between the average price per pack and the number of packs per capita. As the price of cigarette packs rises, fewer people are consuming cigarettes.
- This is to be expected because, as cigarette prices rise, they become less affordable. Therefore, people have limited means to purchase them and instead spend their income on other goods.



Average Price Per Pack vs. Yearly Number Of Packs Per Capita

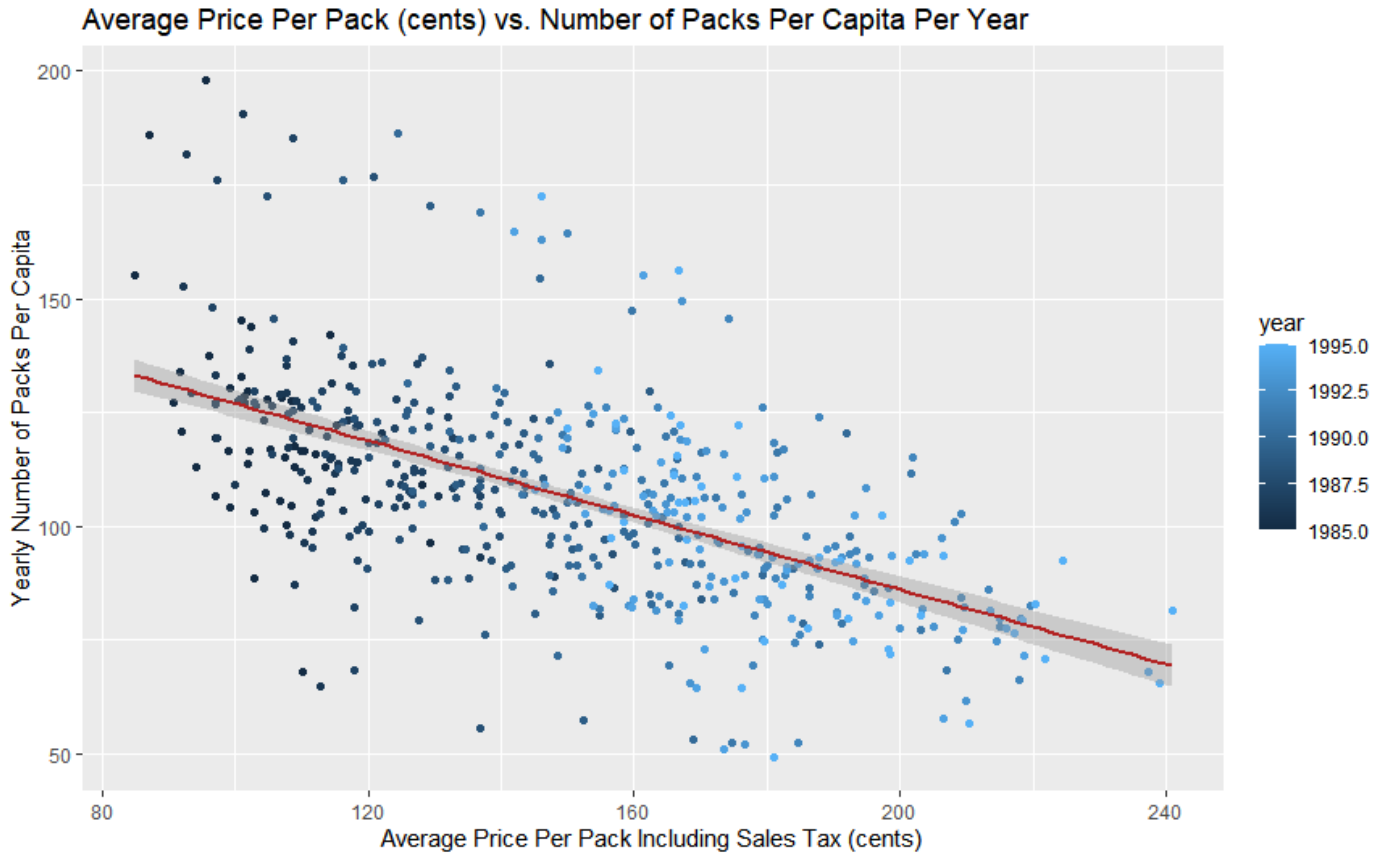
Code to create this scatter plot:

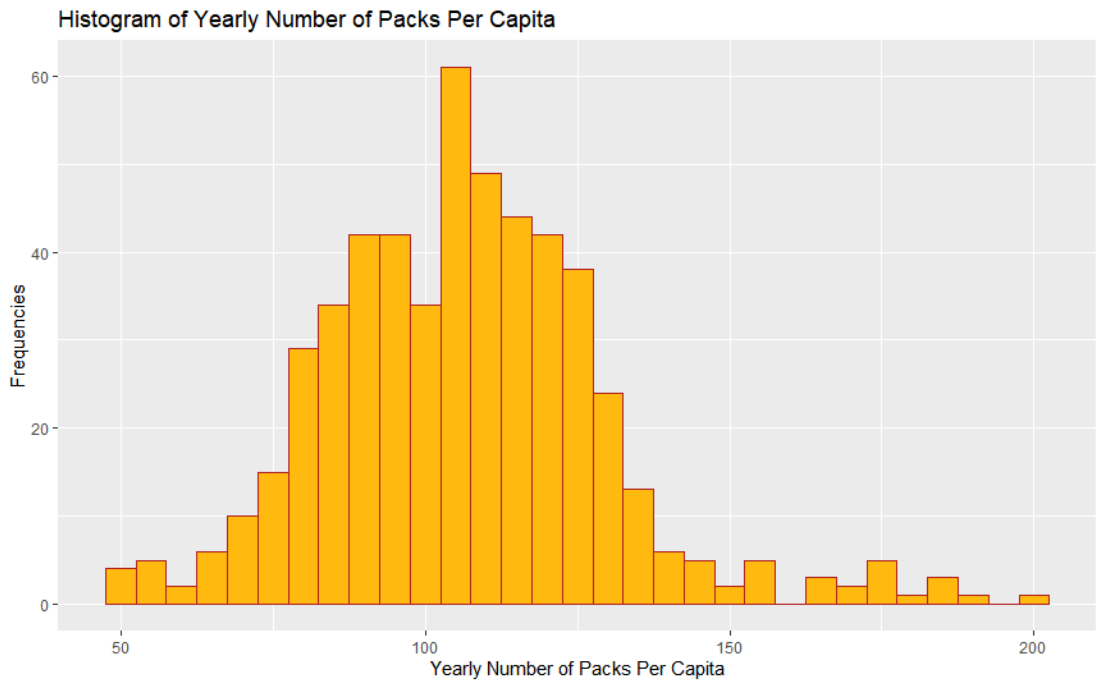
```
ggplot(Cigarette, aes(x = avgprs, y = packpc, color = year)) + geom_point() + geom_smooth(method = lm, color = "firebrick") +
  xlab("Average Price Per Pack Including Sales Tax (cents)") +
  ylab("Yearly Number of Packs Per Capita") +
  ggtitle("Average Price Per Pack (cents) vs. Number of Packs Per Capita Per Year")
```

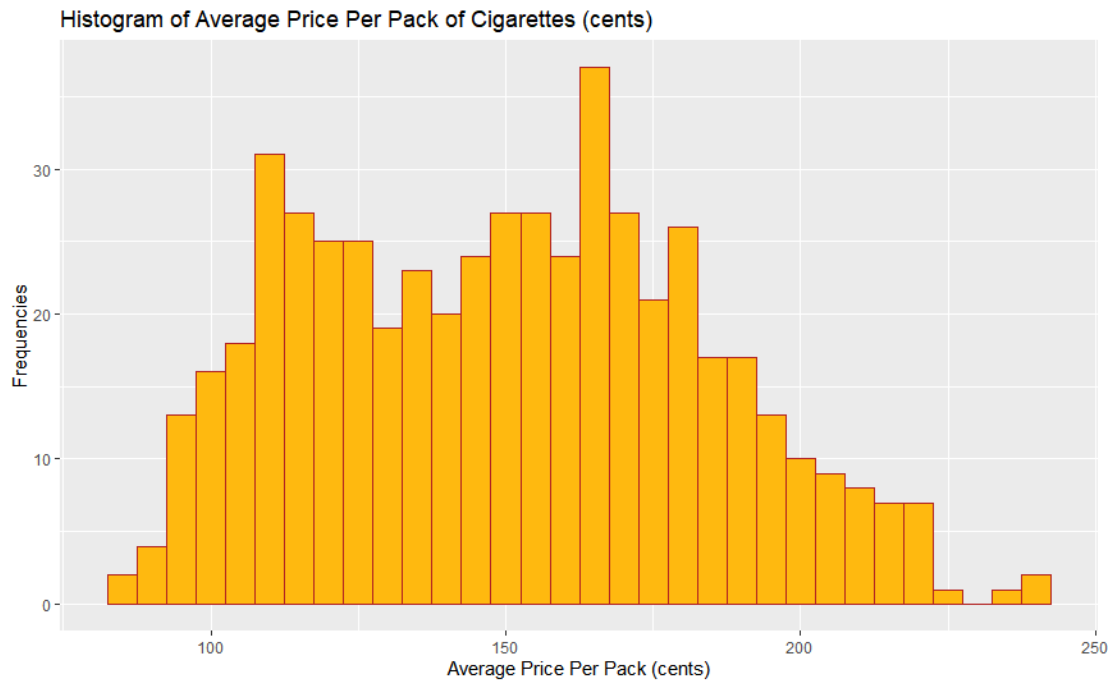


Analysis

- Over the years, the average price per pack has significantly increased.
- Over time, the relationship between the average price per pack and the yearly number of packs per capita appears to become more strongly negative.







Correlation between Average Price Per Pack vs. Yearly Number Of Packs Per Capita

- The yearly average price per pack and the yearly number of packs per capita are both approximately normally distributed.

Correlation between Average Price Per Pack vs. Yearly Number Of Packs Per Capita

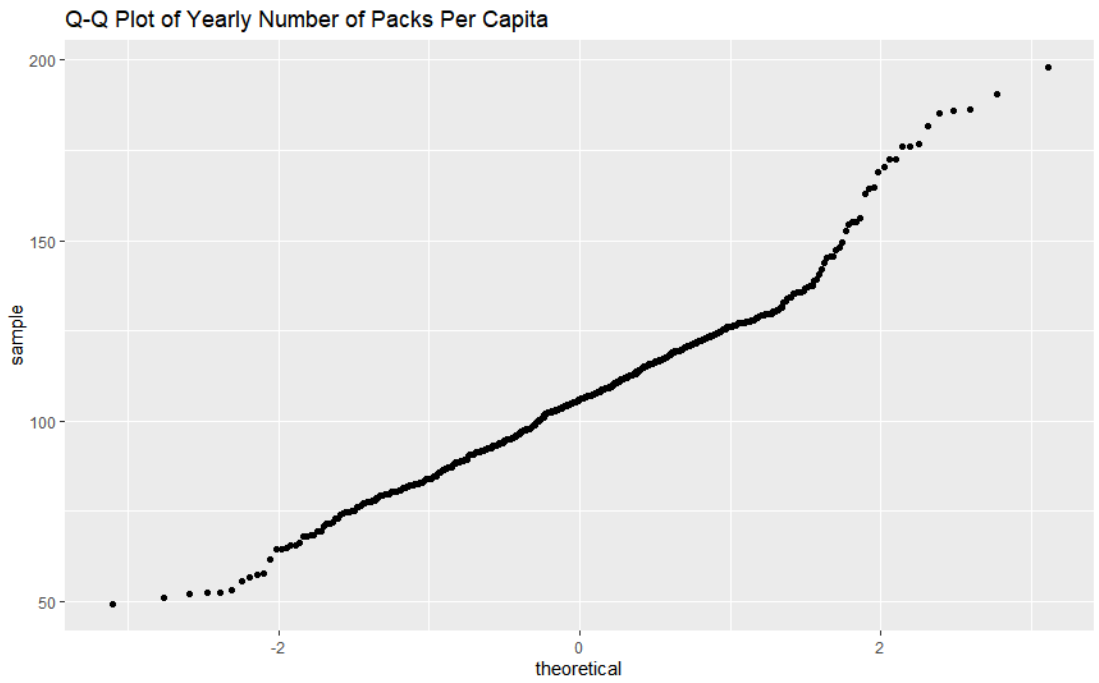
Code to compute both histograms:

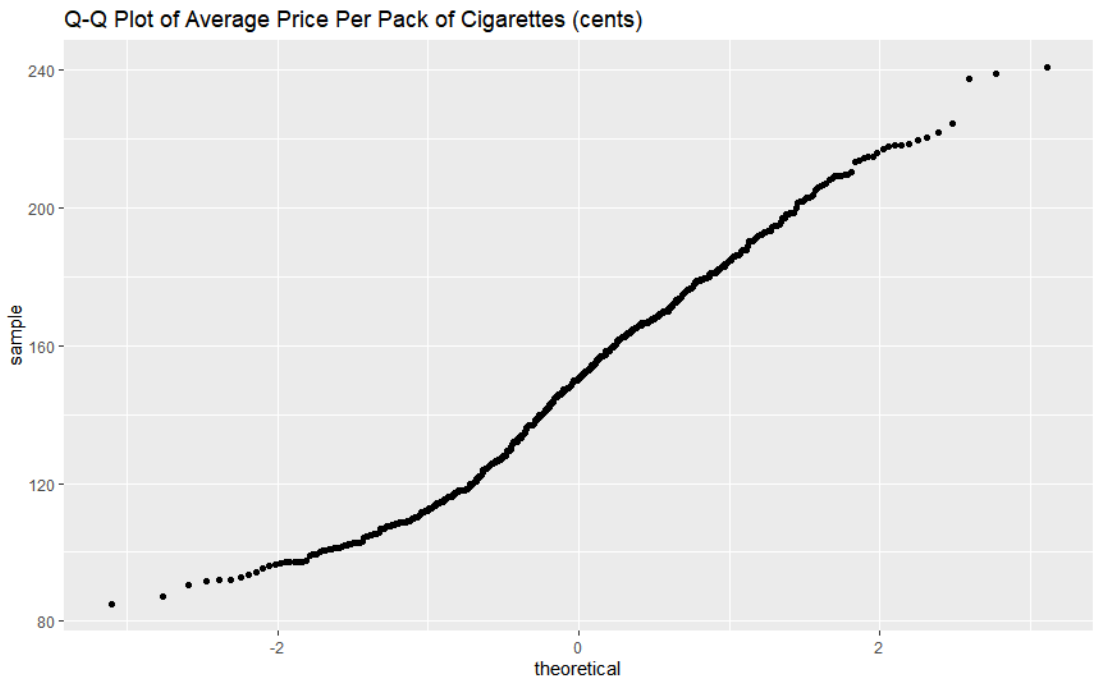
```
avgprs1 <- ggplot(Cigarette, aes(x=avgprs)) +
geom_histogram(binwidth = 5, color = "firebrick", fill
= "darkgoldenrod1") +
  xlab("Average Price Per Pack (cents)") +
ylab("Frequencies") + ggtitle("Histogram of Average
Price Per Pack of Cigarettes (cents)")


packpc1 <- ggplot(Cigarette, aes(x=packpc)) +
geom_histogram(binwidth = 5, color = "firebrick", fill
= "darkgoldenrod1") +
  xlab("Yearly Number of Packs Per Capita") +
ylab("Frequencies") + ggtitle("Histogram of Yearly
Number of Packs Per Capita")


avgprs1

packpc1
```





Correlation between Average Price Per Pack vs. Yearly Number Of Packs Per Capita

- The yearly average price per pack and the yearly number of packs per capita are both approximately normally distributed.

Correlation between Average Price Per Pack vs. Yearly Number Of Packs Per Capita

Code to compute both Q-Q plots:

```
avgprs2 <- ggplot(Cigarette, aes(sample = avgprs)) +
geom_qq() + ggtitle("Q-Q Plot of Average Price Per
Pack of Cigarettes (cents)")
```

```
packpc2 <- ggplot(Cigarette, aes(sample = packpc)) +
geom_qq() + ggtitle("Q-Q Plot of Yearly Number of
Packs Per Capita")
```

```
avgprs2
```

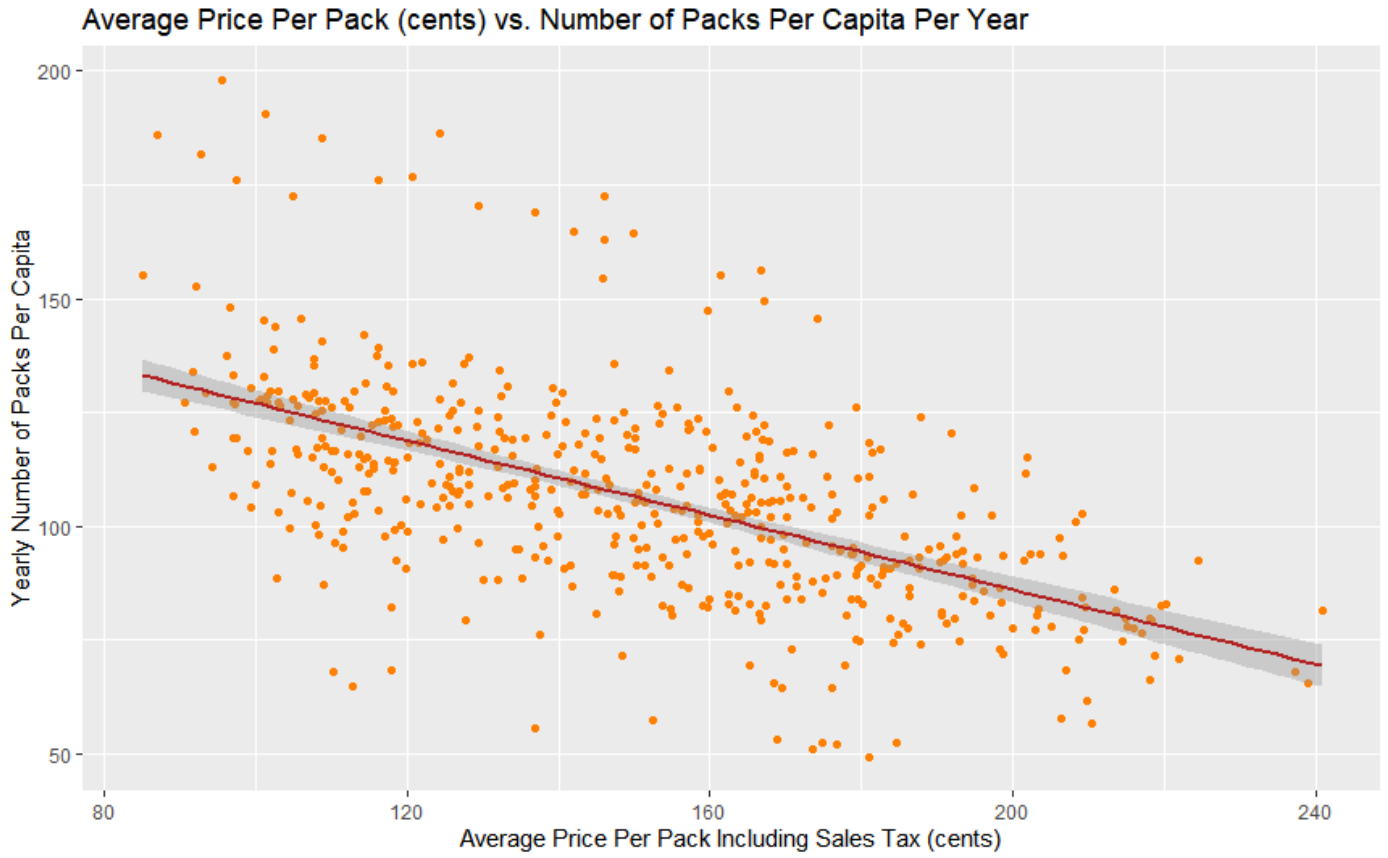
```
packpc2
```

Analysis

- Because both the average price per pack and the yearly number of packs per capita are both normally distributed, the researcher used Pearson's product-moment correlation to compute r, the correlation coefficient.
- The correlation coefficient, r, is -0.5854443. This means that the correlation between the average price per pack and the number of packs per capita per year is **moderately** negative.

Code to compute correlation strength:

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



Analysis

- The linear regression for the Average Price Per Pack (cents) vs. Number of Packs Per Capita Per Year is:

y = -0.4088 x + 167.8774

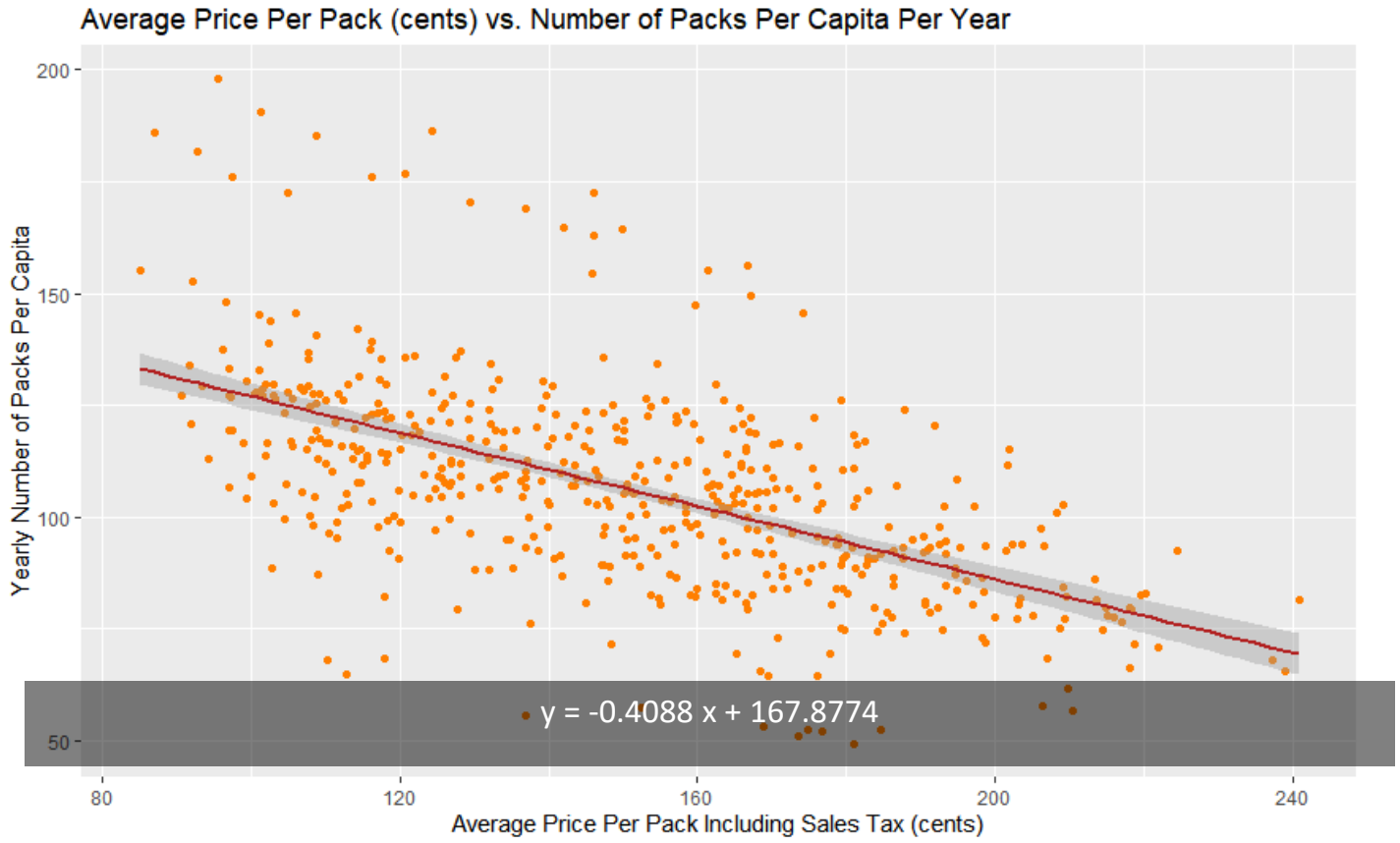
x = Average Price Per Pack (cents)

y = Number of Packs Per Capita Per Year

Code to compute this linear regression:

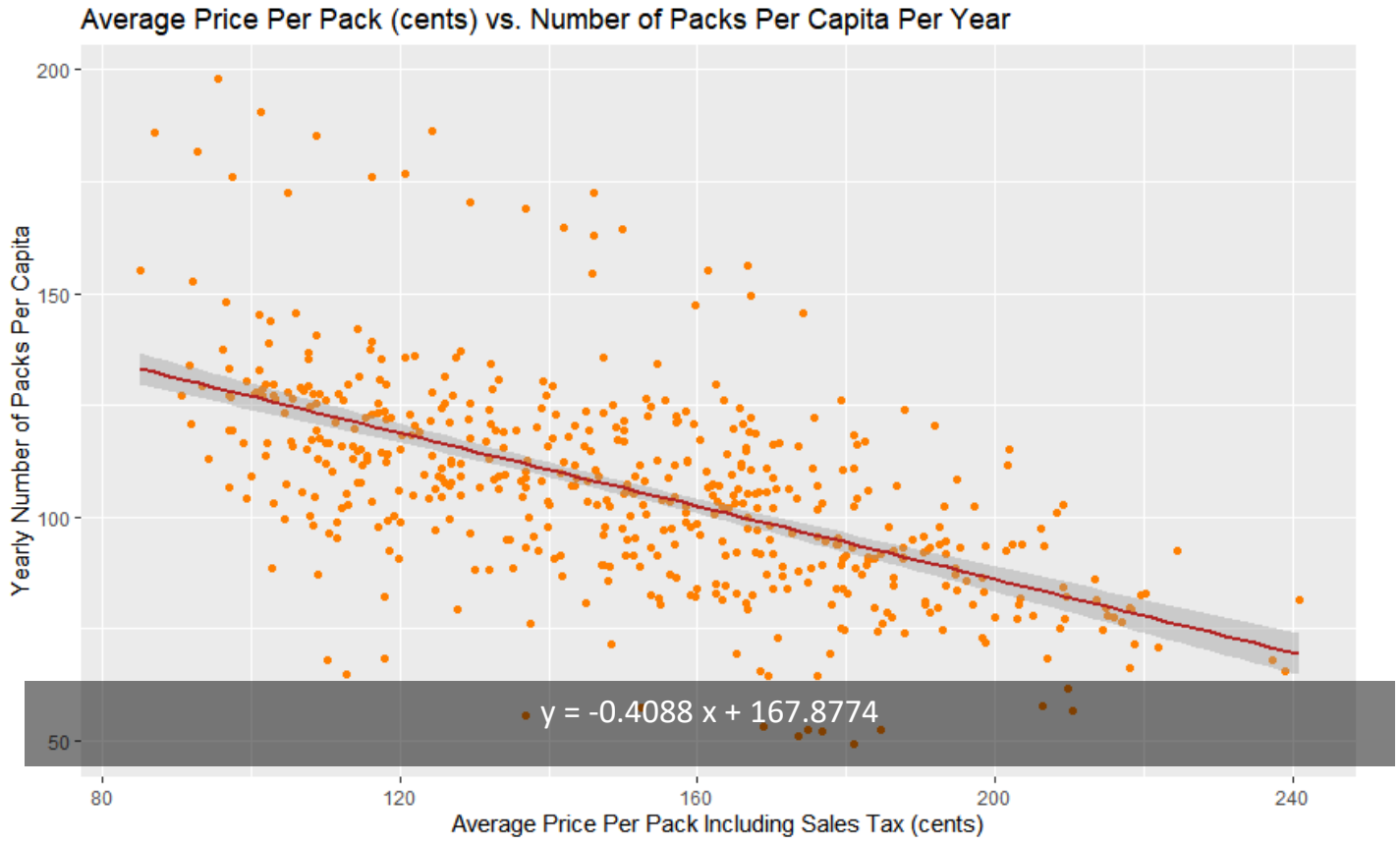
regression1 <- lm(packpc ~ avgprs, Cigarette)

regression1



Analysis

- The linear regression indicates that for every 100-cent increase in the price per pack, the yearly number packs per capita would decrease by approximately 40.

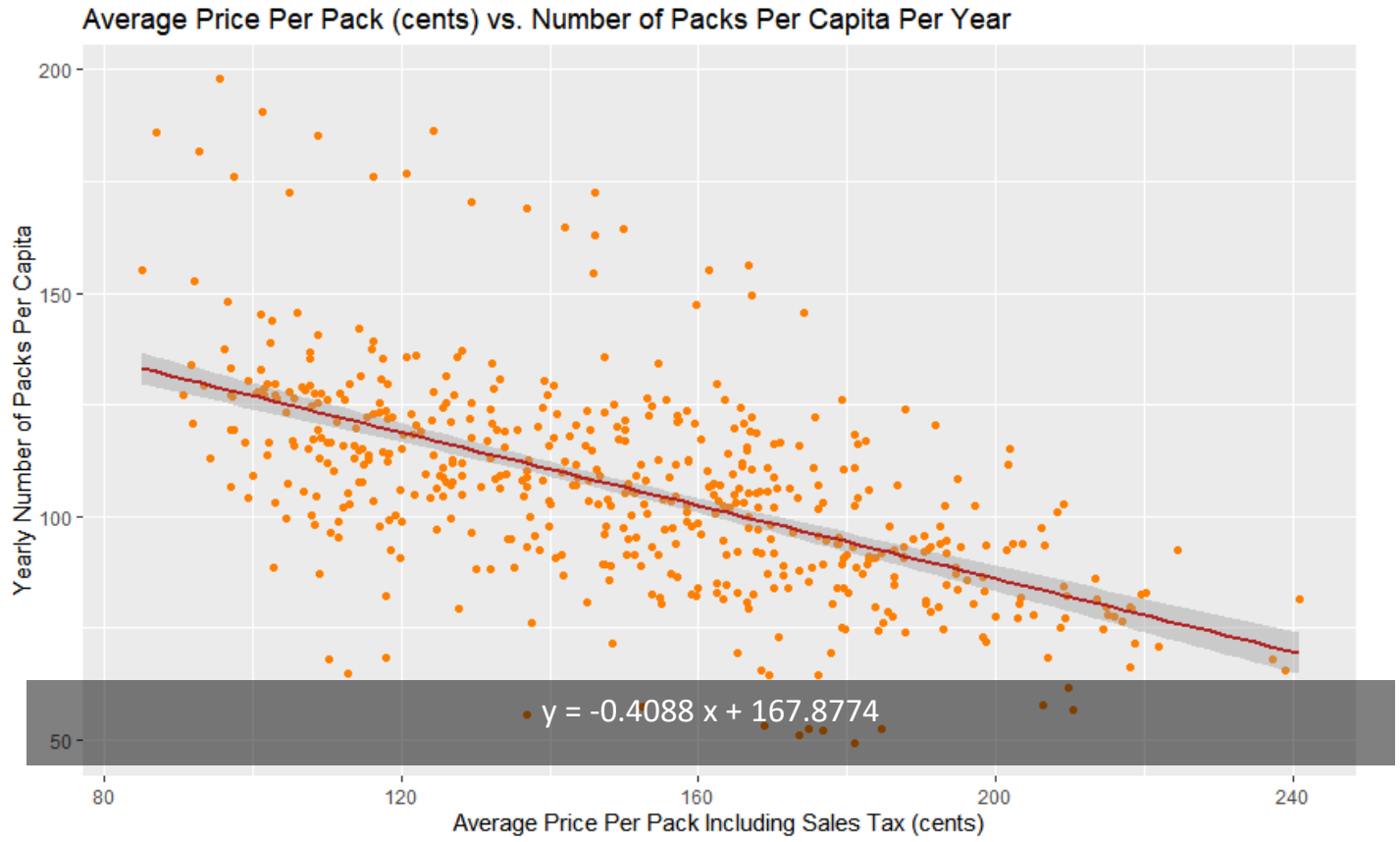


Analysis

- The p-value of the average price per pack is 2e-16, which is well below 0.05 and 0.001. Therefore, the regression is statistically significant.
- The adjusted r-squared value is 0.3415. This means that linear regression explains approximately 34.15% of the variability of the relationship between these two variables.
- The confidence interval narrows at a price of 160 cents per pack and widens and the upper and lower price extremes between 1985 – 1995. Therefore, the accuracy of the linear regression improves near the median average price per pack.

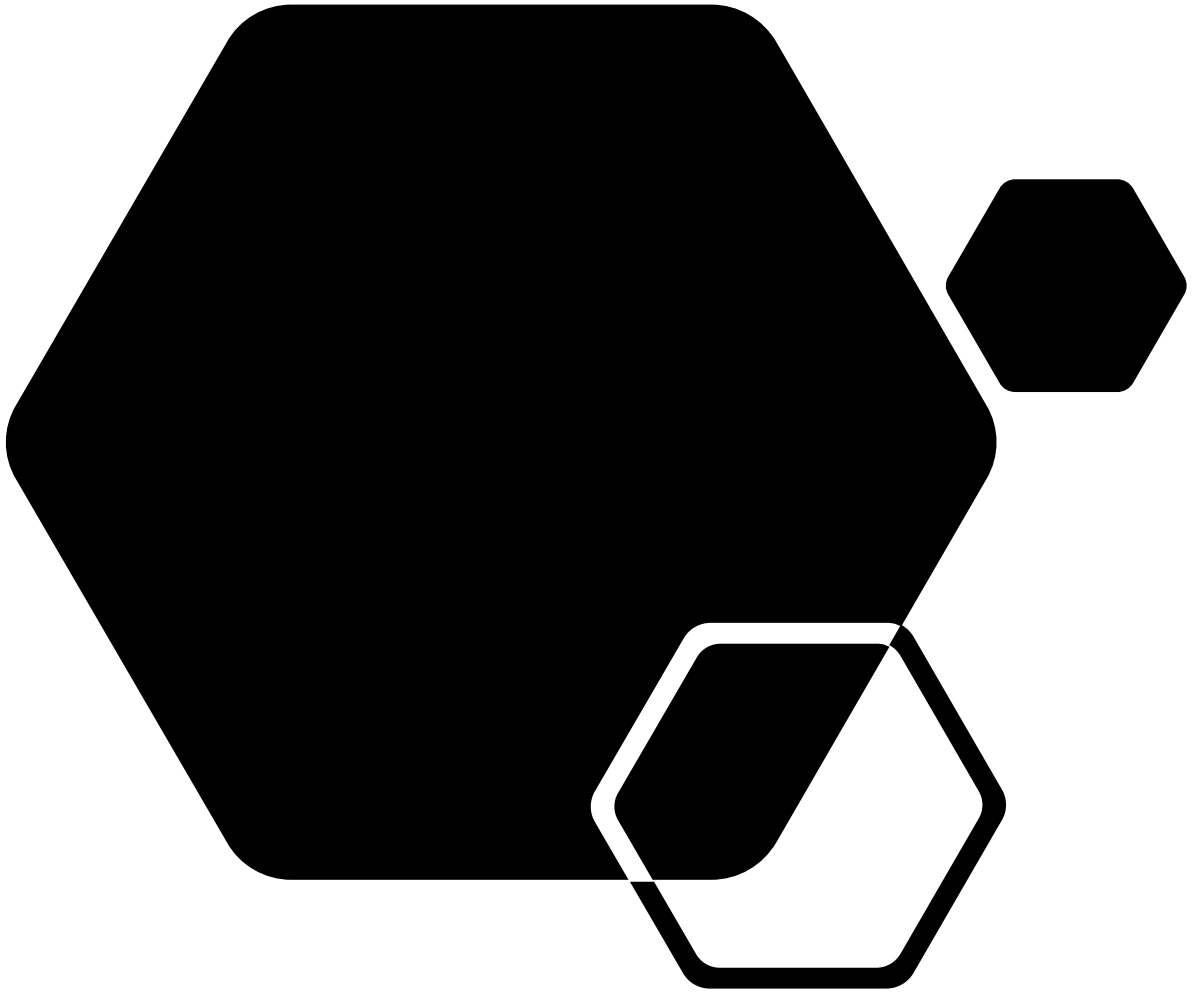
Code to compute summary statistics of linear regression:

summary(regression1)



Part 4

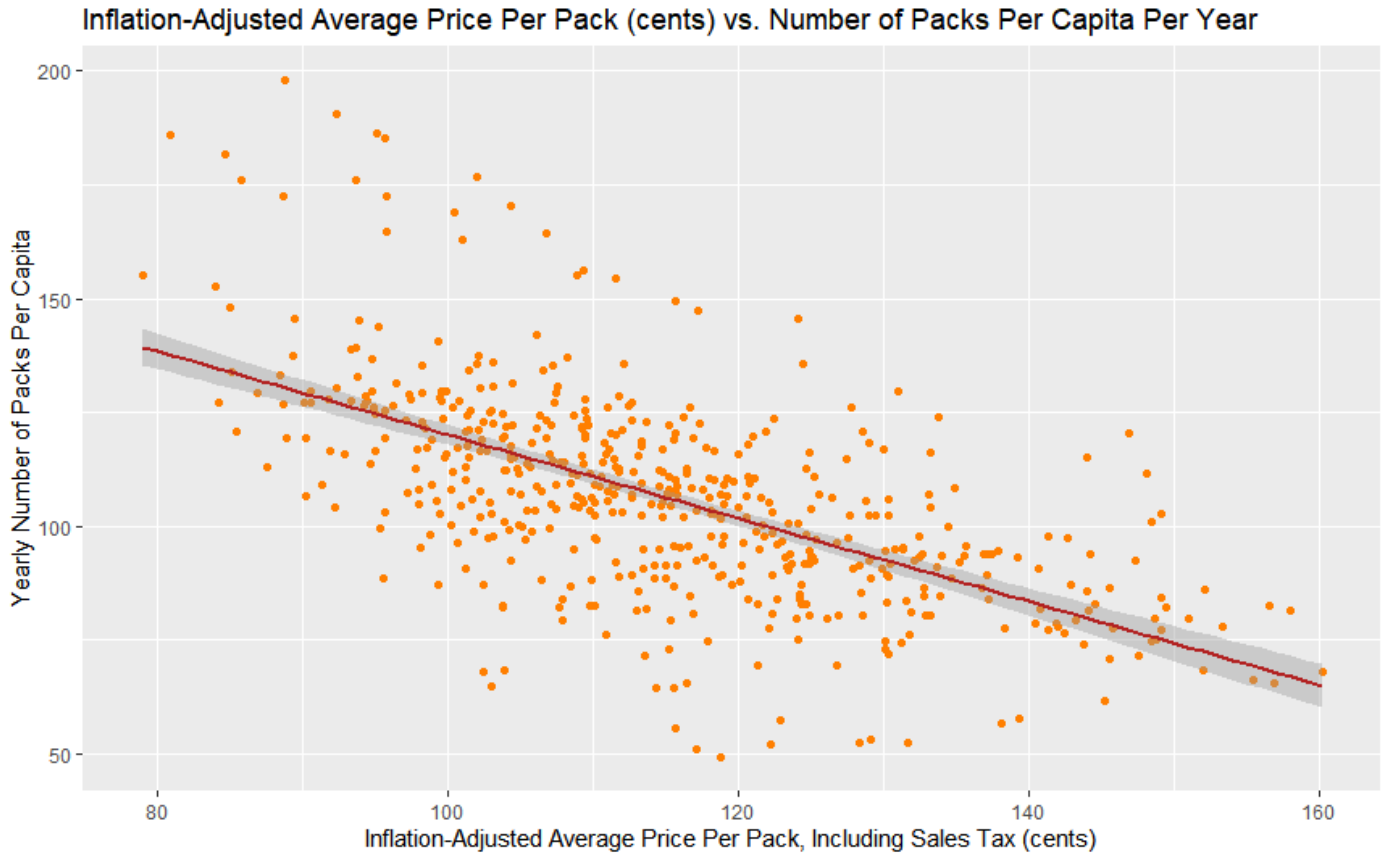
Inflation-Adjusted Average Price Per Pack vs. Yearly Number Of Packs Per Capita



Inflation-Adjusted Average Price Per Pack vs. Yearly Number Of Packs Per Capita

Code to create this scatter plot:

```
ggplot(Adjusted, aes(x = Adj, y = packpc)) +
geom_point(color = "darkorange1") +
geom_smooth(method = lm, color = "firebrick")
+
  xlab("Inflation-Adjusted Average Price Per
Pack, Including Sales Tax (cents)") +
  ylab("Yearly Number of Packs Per Capita") +
  ggtitle("Inflation-Adjusted Average Price Per
Pack (cents) vs. Number of Packs Per Capita Per
Year")
```



Analysis

- The linear regression for the Inflation-Adjusted Average Price Per Pack (cents) vs. Number of Packs Per Capita Per Year is:

y = -0.9164 x + 211.7682

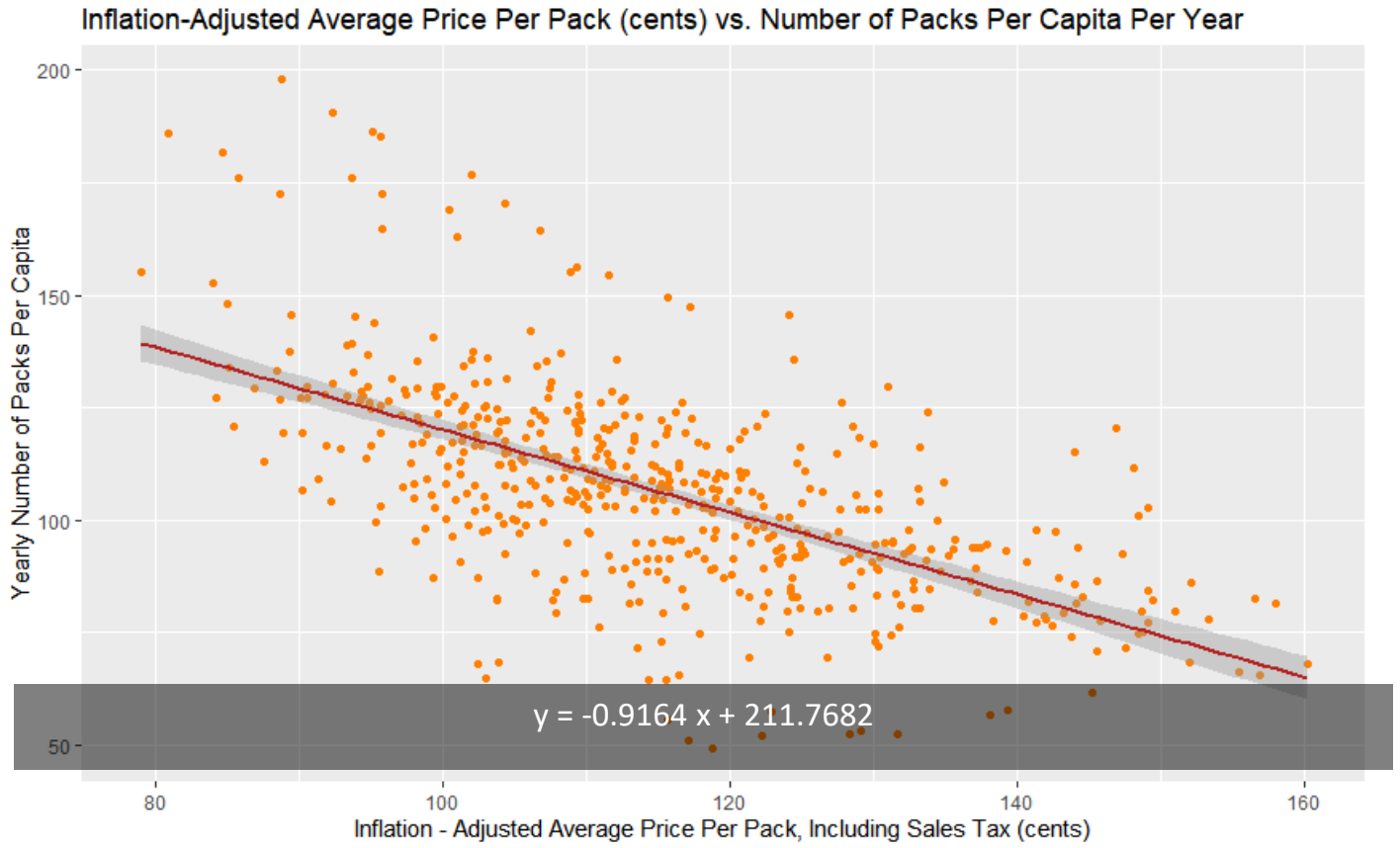
x = Inflation-Adjusted Average Price Per Pack (cents)

y = Number of Packs Per Capita Per Year

Code to compute this linear regression:

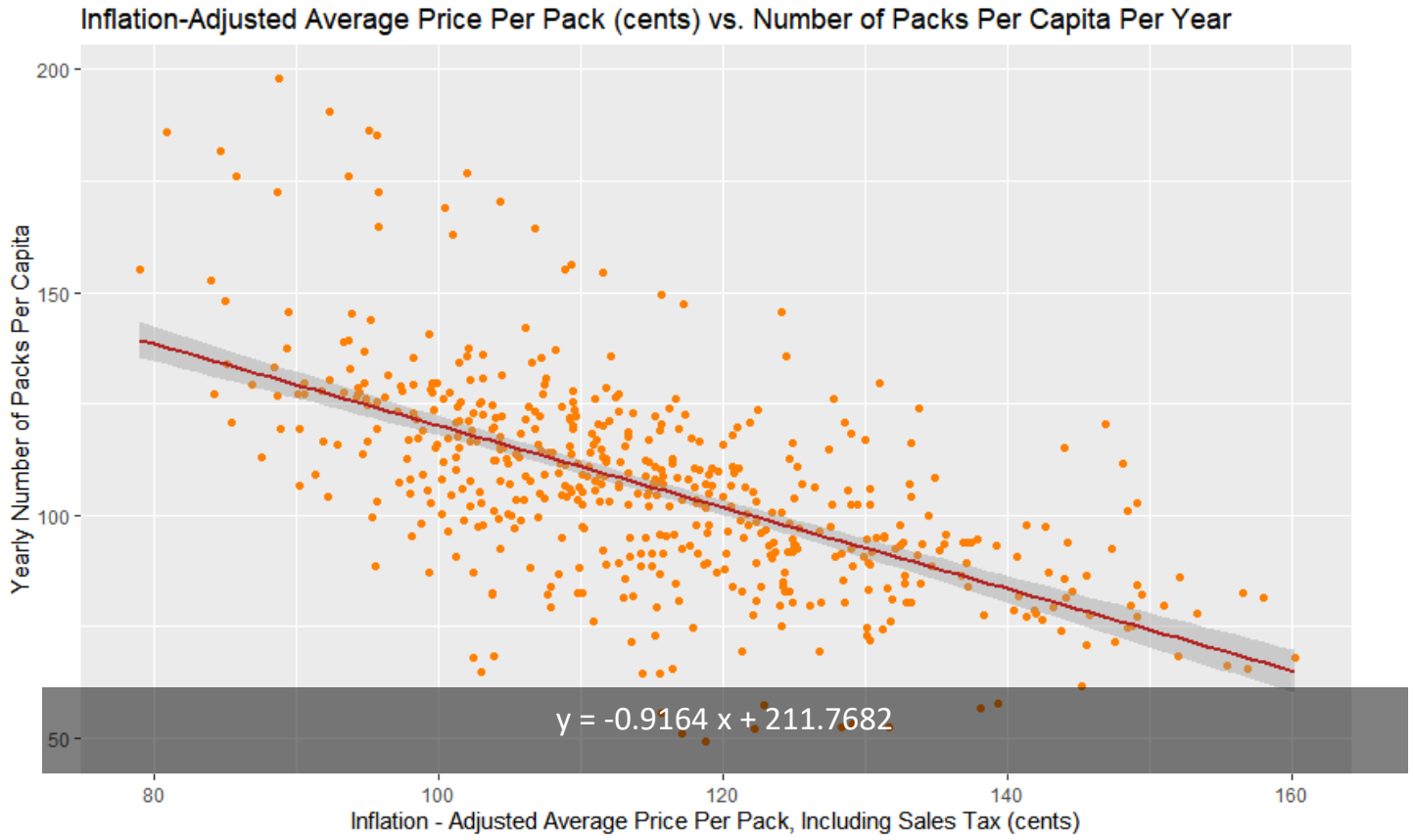
```
Adjusted <- Cigarette %>% mutate(Adj = (avgprs/cpi))
View(Adjusted)

regression2 <- lm(packpc ~ Adj, Adjusted)
regression2
```



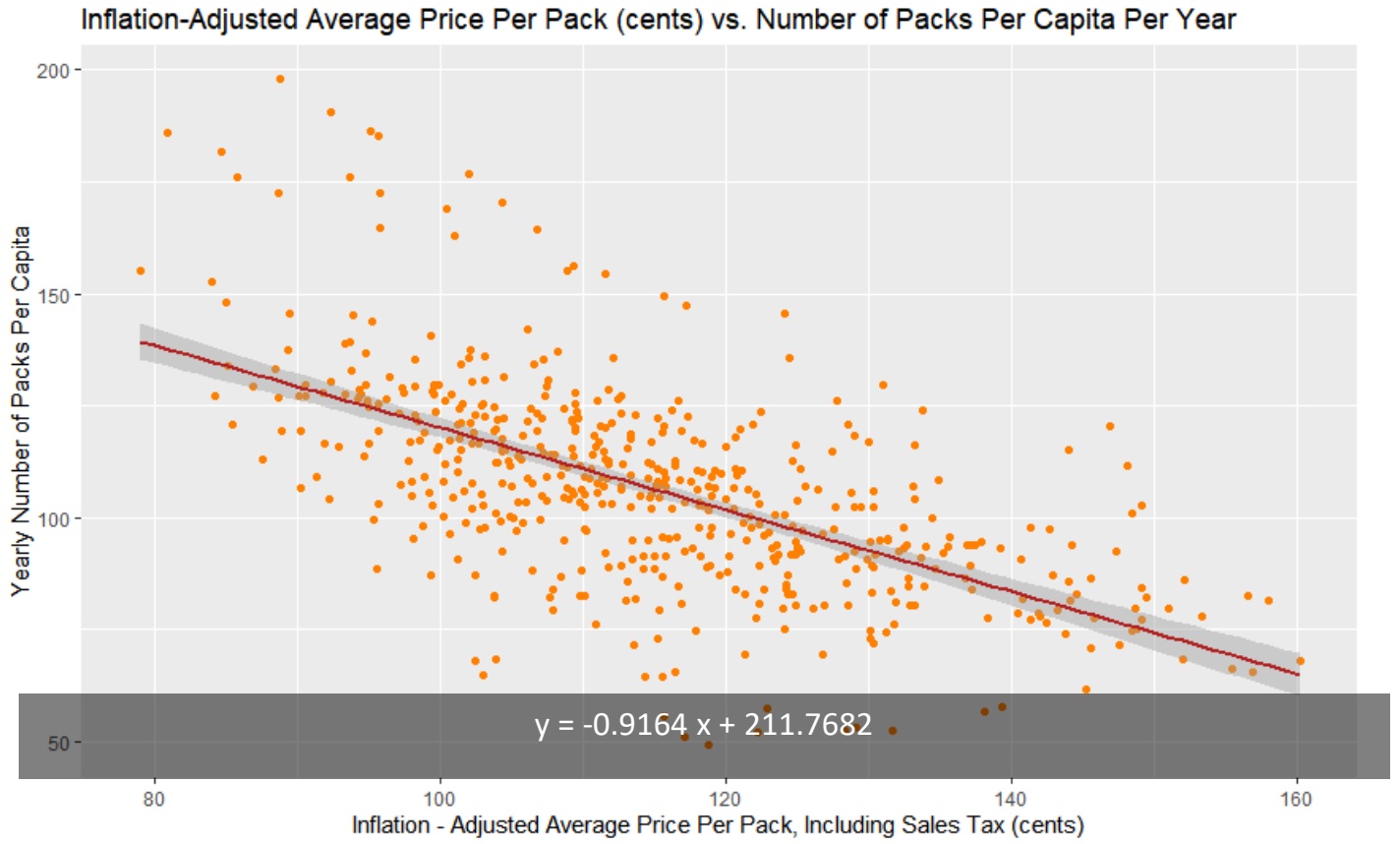
Analysis

- After adjusting the average price of a pack of cigarettes for inflation, the correlation between the cost of cigarettes and their consumption is even more negative. For every increase in 100 cents in the price per pack, the yearly number per capita drops by approximately 92.
- Before the inflation adjustment, the yearly number packs per capita would decrease by approximately 40.



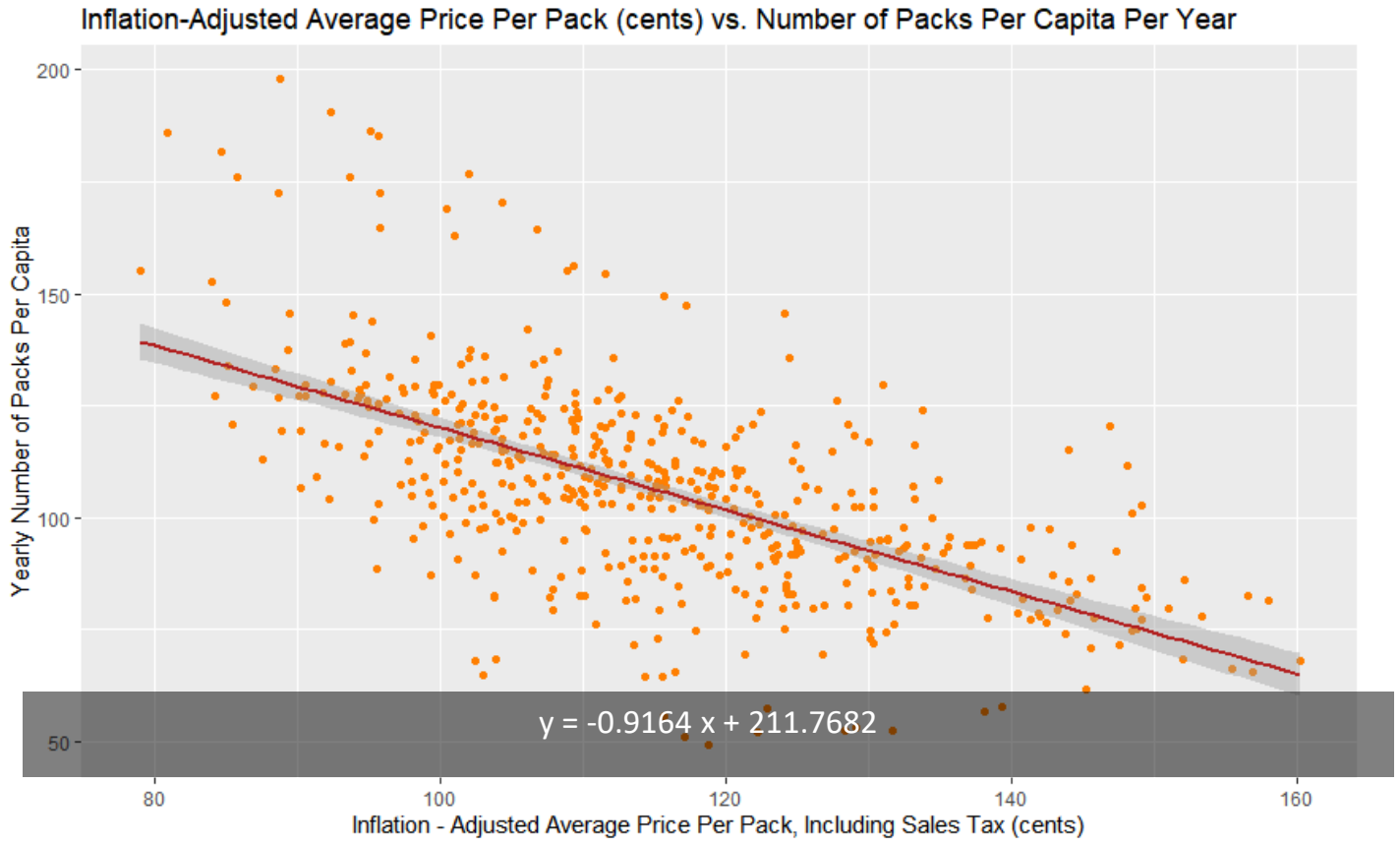
Analysis

- After adjusting the average price of a pack of cigarettes for inflation, the correlation between the cost of cigarettes and their consumption is even more negative. For every increase in 100 cents in the price per pack, the yearly number per capita drops by approximately 92.
- Before the inflation adjustment, the yearly number packs per capita would decrease by approximately 40.



Analysis

- After adjusting the average price of a pack of cigarettes for inflation, the correlation coefficient, r, is -0.6138902.
- Before the inflation adjustment, the correlation coefficient, r, was -0.5854443.
- Despite the increased strength of the correlation between the cost and prevalence of cigarettes, the correlation remains moderately negative after the inflation adjustment.

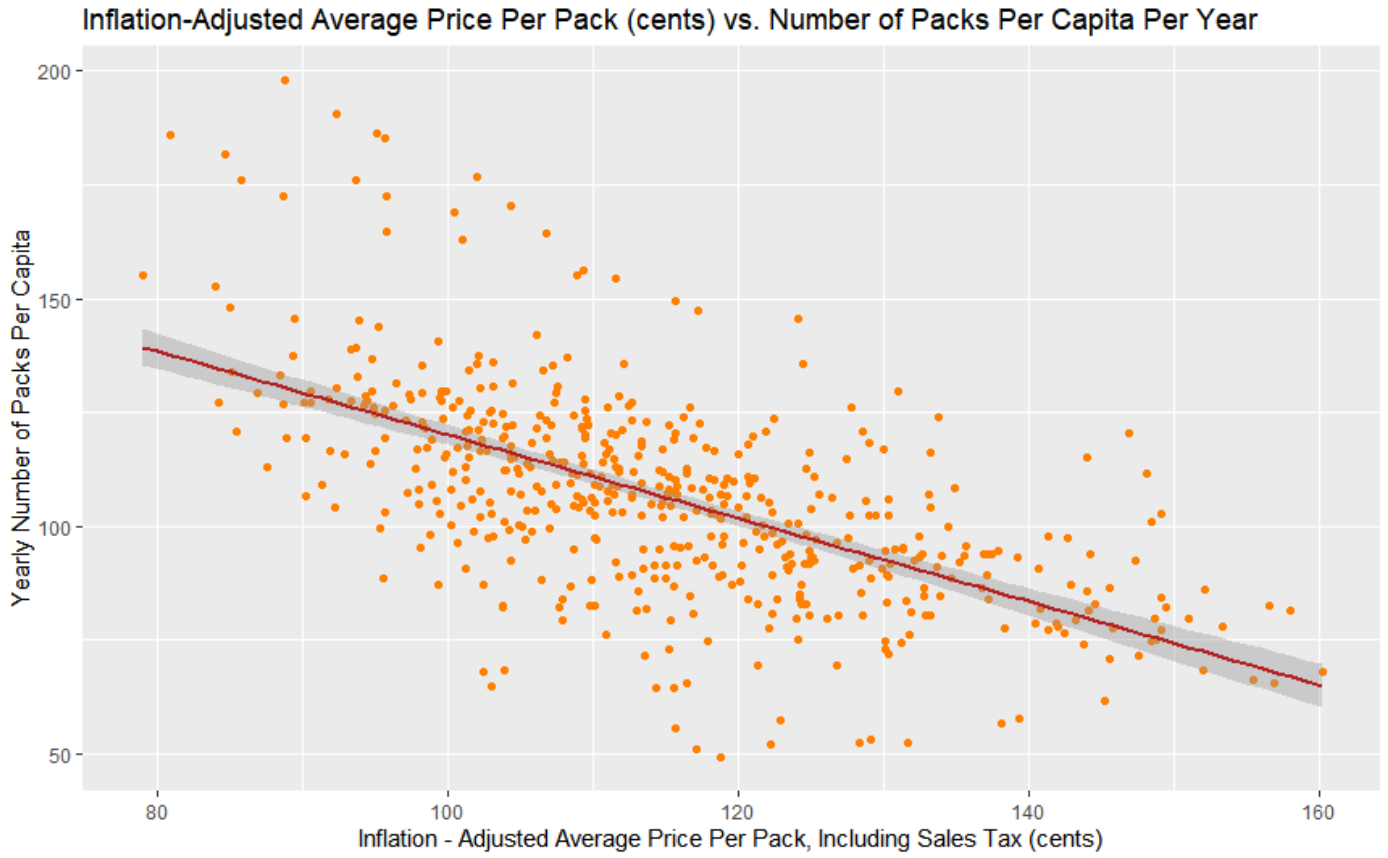


Analysis

- After the inflation adjustment, the p-value of the average price per pack remains 2e-16, which is well below 0.05 and 0.001. Therefore, the regression continues to be statistically significant.
- The adjusted r-squared value is 0.3757. This means that linear regression explains approximately 35.57% of the variability of the relationship between these two variables.
- The confidence interval narrows at a price of 120 cents per pack and widens and the upper and lower price extremes between 1985 – 1995. Therefore, the accuracy of the linear regression improves near the median average price per pack.

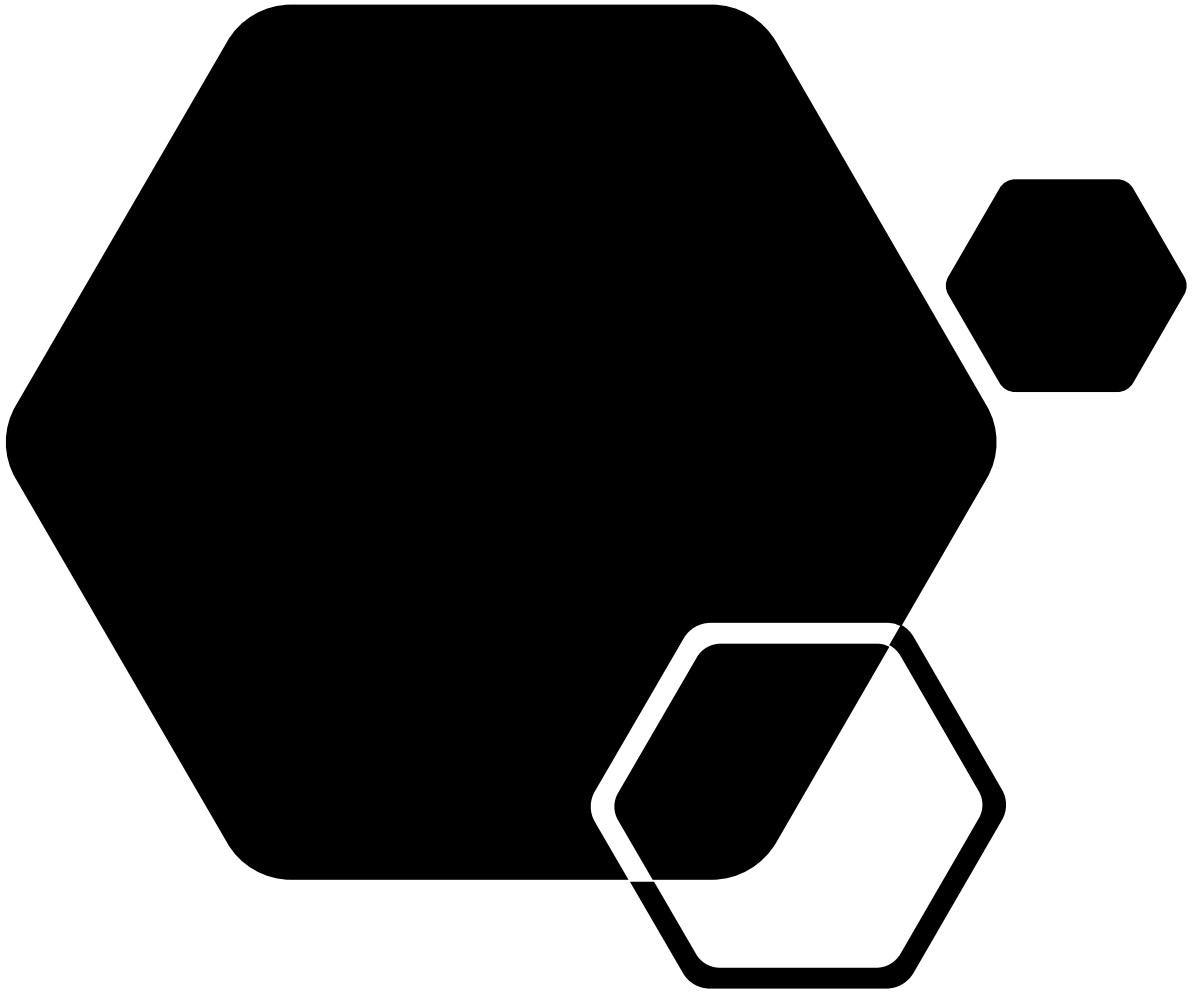
Code to compute summary statistics of linear regression:

summary(regression2)



Part 5

Number of packs per capita in 1985 vs. that of 1995



Number of packs per capita in 1985 vs. that of 1995

Code to implement paired t-test:

```
Cigs85 <- Cigarette %>% select(year, packpc) %>% filter(year == 1985)
View(Cigs85)
Cigs95 <- Cigarette %>% select(year, packpc) %>% filter(year == 1995)
View(Cigs95)

t.test(Cigs85$packpc, Cigs95$packpc, paired = TRUE)
```

- Results of paired t-test:
  - t = 14.789
  - df = 47
  - p-value < 2.2e-16
  - mean of the differences = 25.70863

Number of packs per capita in 1985 vs. that of 1995

- Results of paired t-test:
  - t = 14.789
  - df = 47
  - p-value < 2.2e-16
  - mean of the differences = 25.70863
- A paired, or dependent, t-test, is a test which compares the means of two related groups to determine whether there is a statistically significant difference between these means.
- In this case, the packs per capita in 1985 and 1995, respectively, are related by time. This t-test would determine whether there is a change in cigarette consumption between 1985 and 1995.

Number of packs per capita in 1985 vs. that of 1995

- Results of paired t-test:
  - t = 14.789
  - df = 47
  - p-value < 2.2e-16
  - mean of the differences = 25.70863
- The t-value is 14.789. The t-value is size of the difference relative to the variation in the sample data. A t-value of 0 would mean that there is no discernible difference between the cigarette consumption in 1985 versus that of 1995.

Number of packs per capita in 1985 vs. that of 1995

- Results of paired t-test:
  - t = 14.789
  - df = 47
  - p-value < 2.2e-16
  - mean of the differences = 25.70863
- Because there are 48 observations in both data sets, there are 47 degrees of freedom (df).
- Once df > 35, the t-distribution is said to approximate the normal distribution.

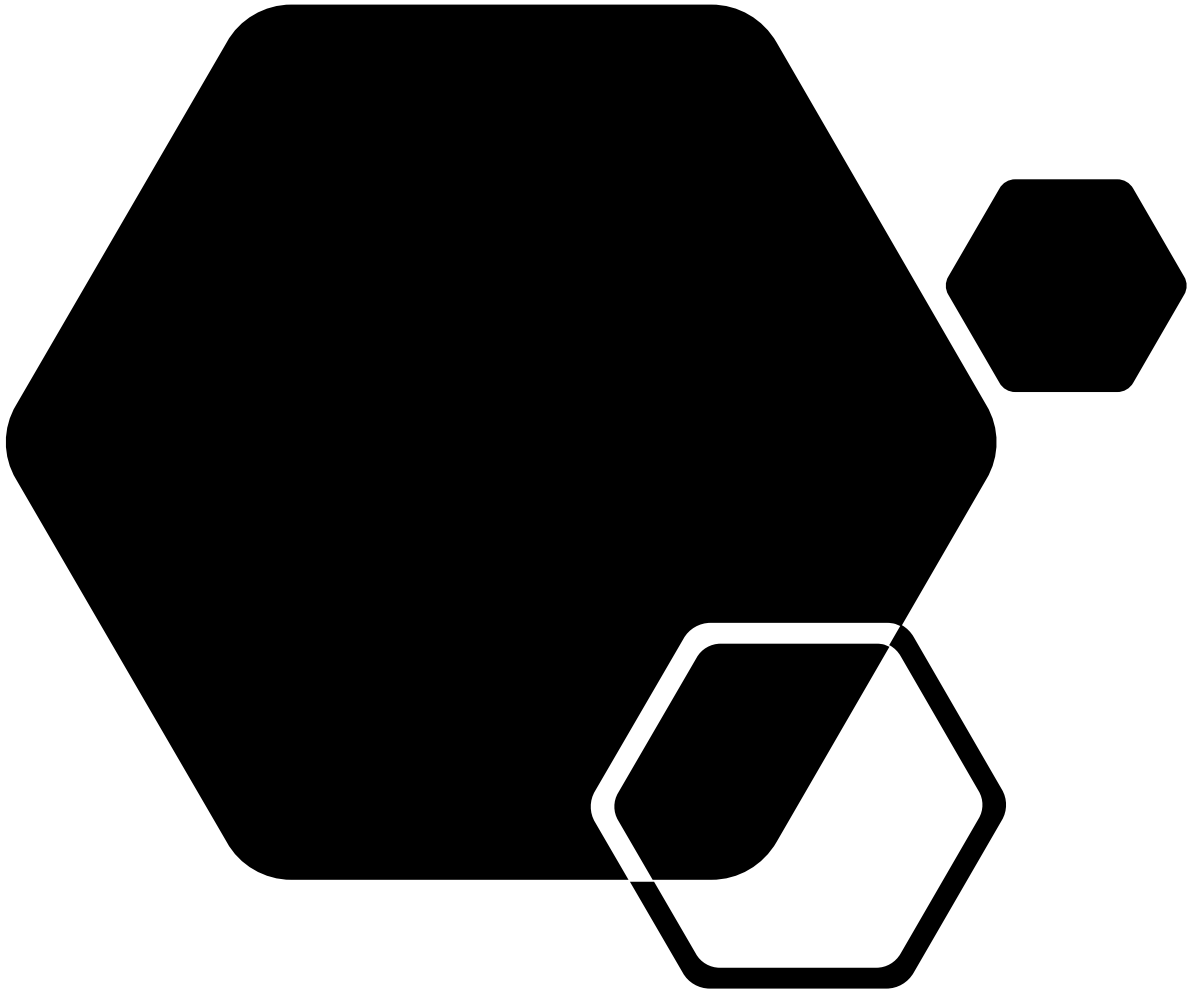
Number of packs per capita in 1985 vs. that of 1995

- Results of paired t-test:
  - t = 14.789
  - df = 47
  - p-value < 2.2e-16
  - mean of the differences = 25.70863

- A p-value is the probability of obtaining test results at least as extreme as the actual results, under the assumption that the null hypothesis is correct.
- The results show that the p-value, $p < 2.2e\text{-}16 < 0.05$. Therefore, there is sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. This means that there is a statistically significant difference in cigarette consumption between 1985 and 1995.
- The mean of the differences in cigarette consumption between 1985 and 1995 is 25.70863 . This means that, in 1995, people have been purchasing 25 packs of cigarettes less than in 1985.

Part 6

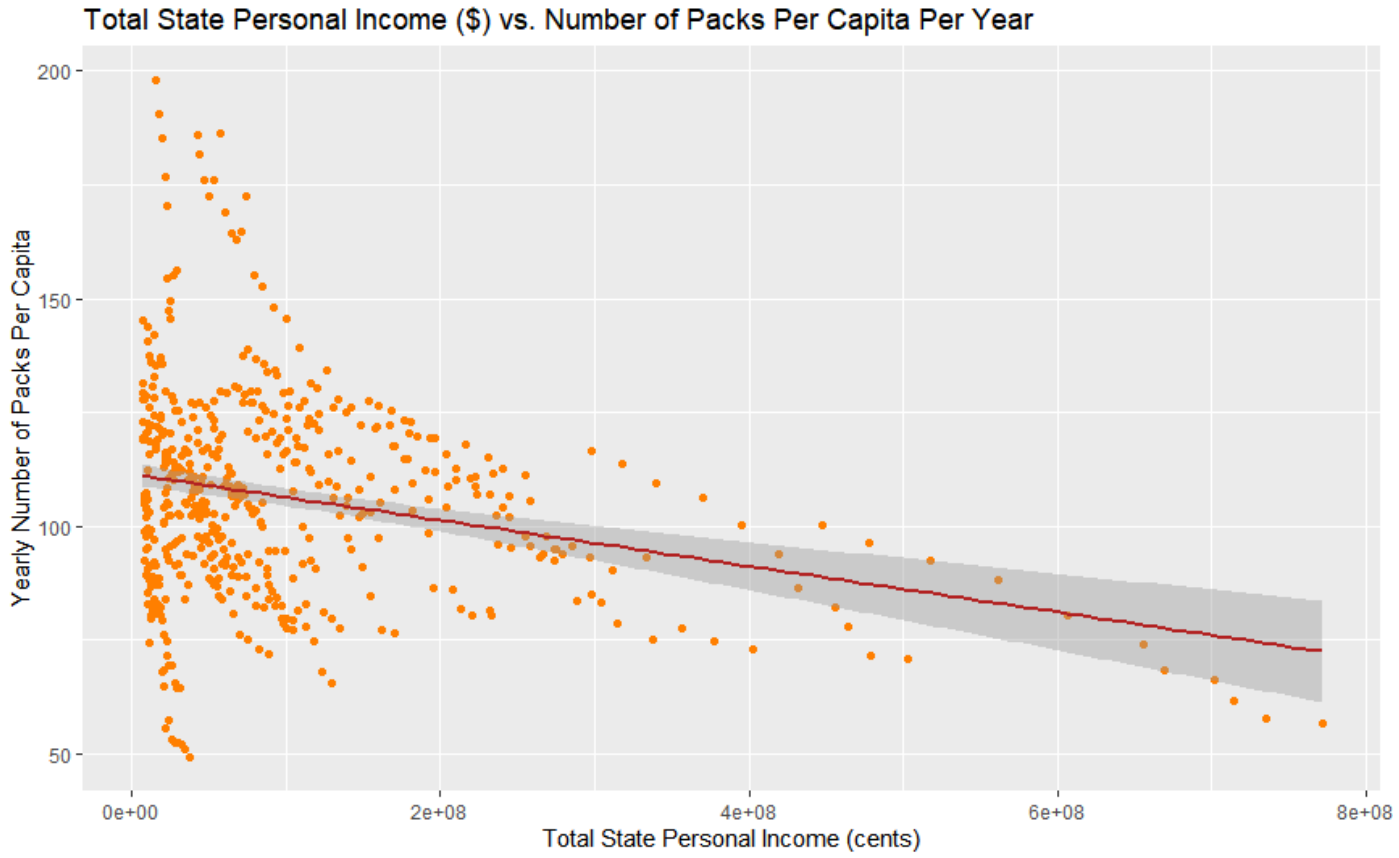
Packs Per Capita vs. Income



Packs Per Capita vs. Income

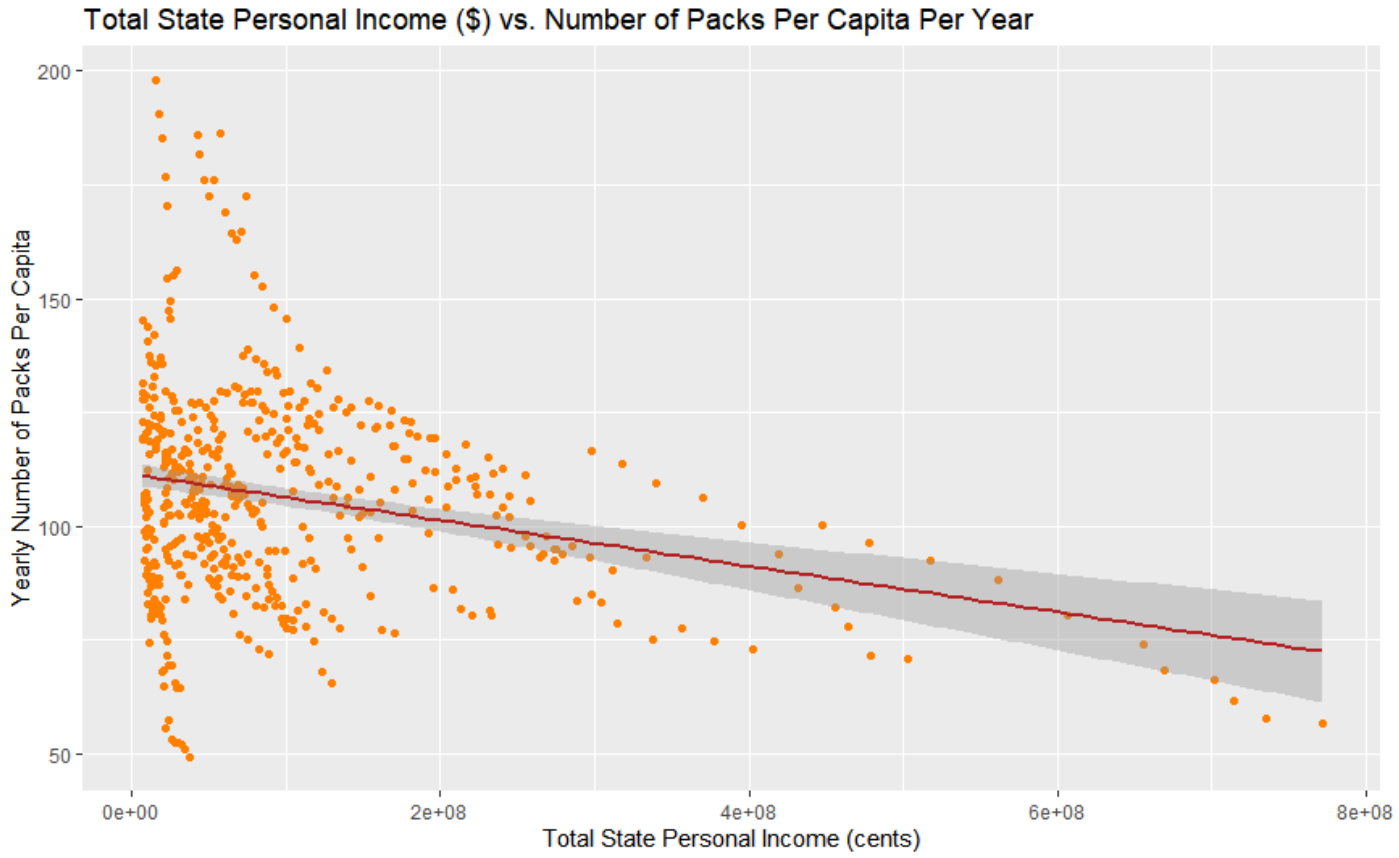
Code to create this scatter plot:

```
ggplot(Cigarette, aes(x = income, y = packpc)) +
geom_point(color = "darkorange1") +
  geom_smooth(method = lm, color =
"firebrick") +
  xlab("Total State Personal Income (cents)") +
  ylab("Yearly Number of Packs Per Capita") +
  ggtitle("Total State Personal Income ($) vs.
Number of Packs Per Capita Per Year")
```



Analysis

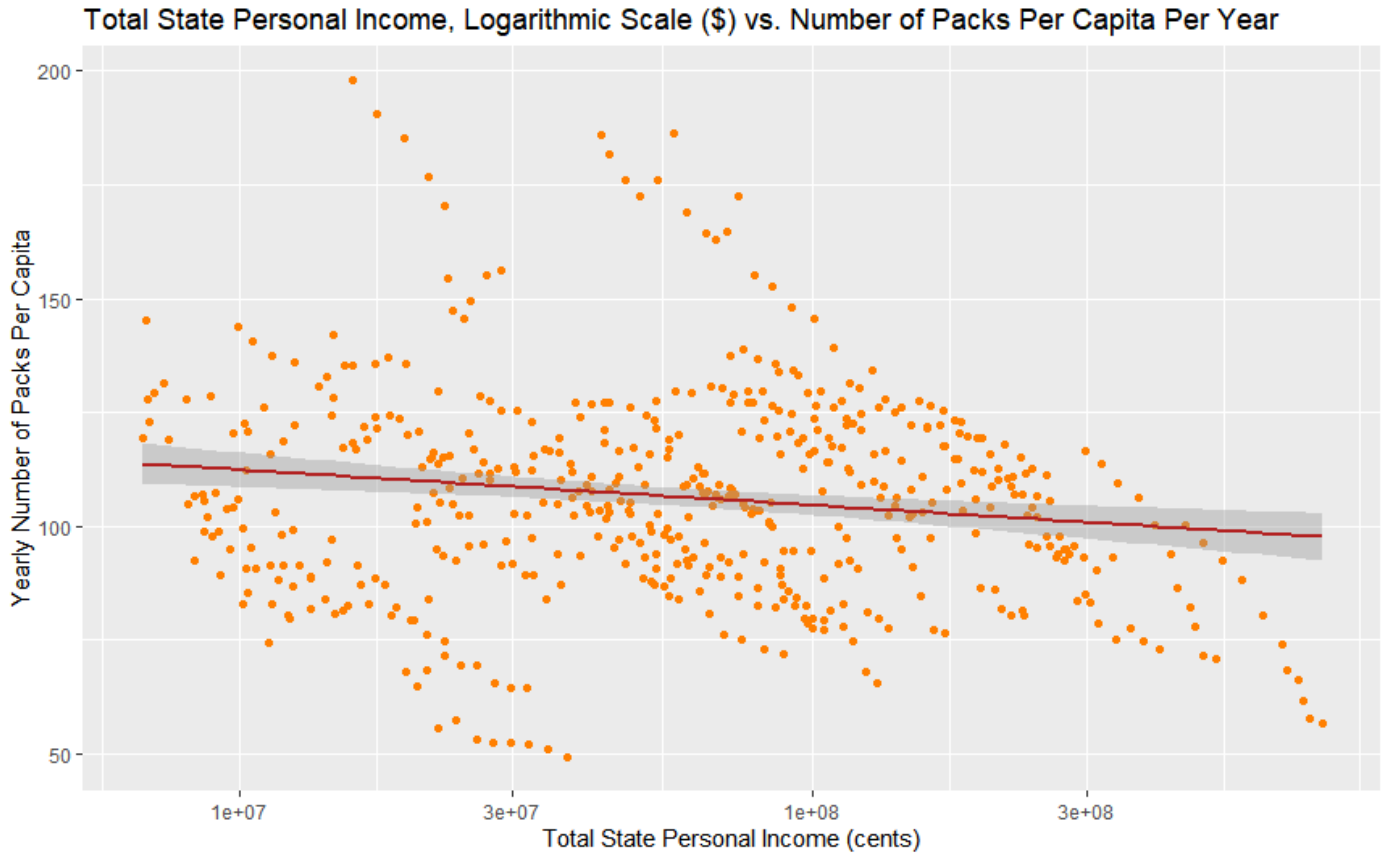
- This scatter plot shows that states with lower total personal income have increased prevalence in cigarette consumption.
- According to the CDC, tobacco companies tend to market their products more in lower income neighborhoods and states. There is also a higher density of tobacco retailers in lower-income and minority neighborhoods.



Packs Per Capita vs. Income (Logarithmic Scale)

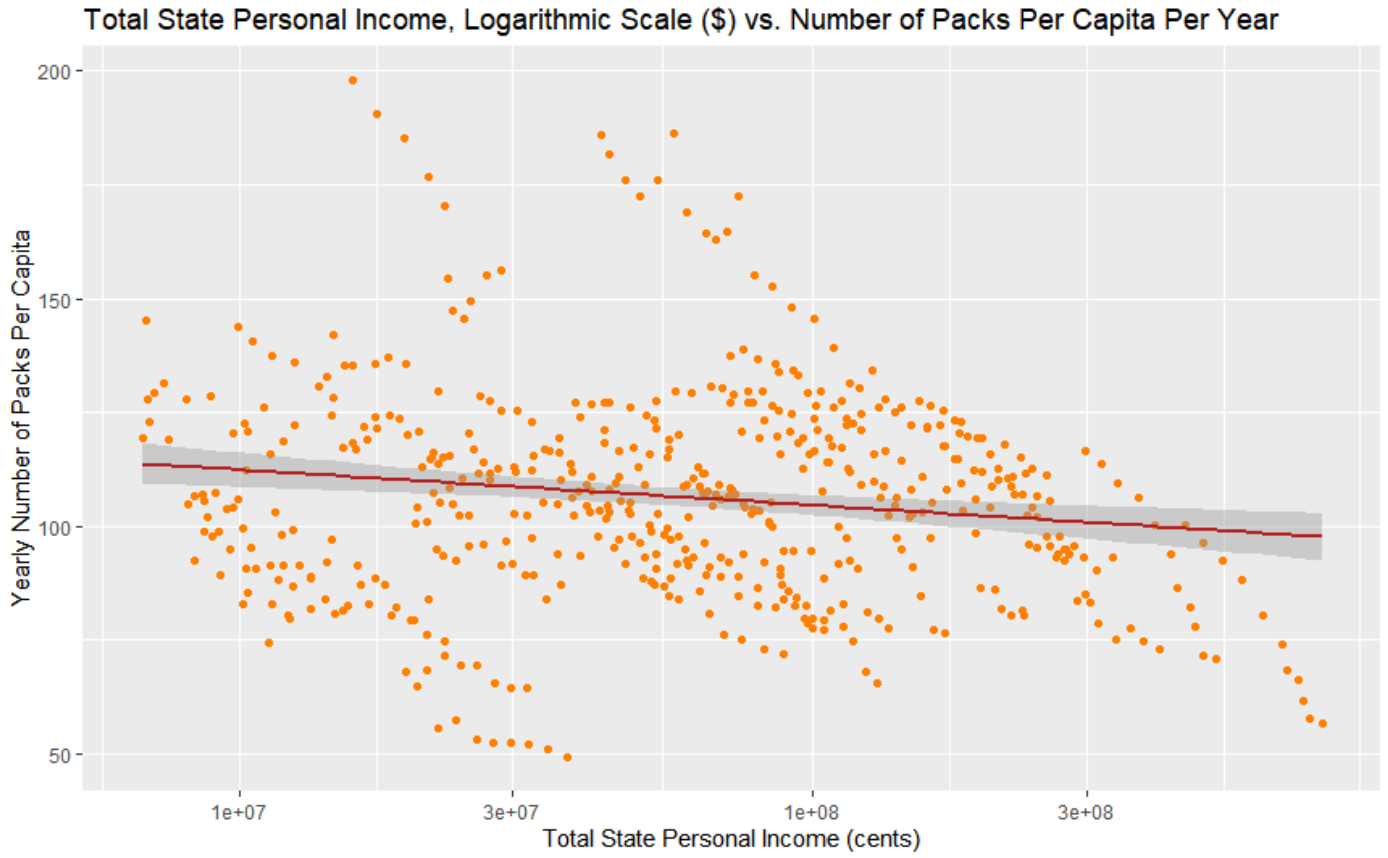
Code to create this scatter plot:

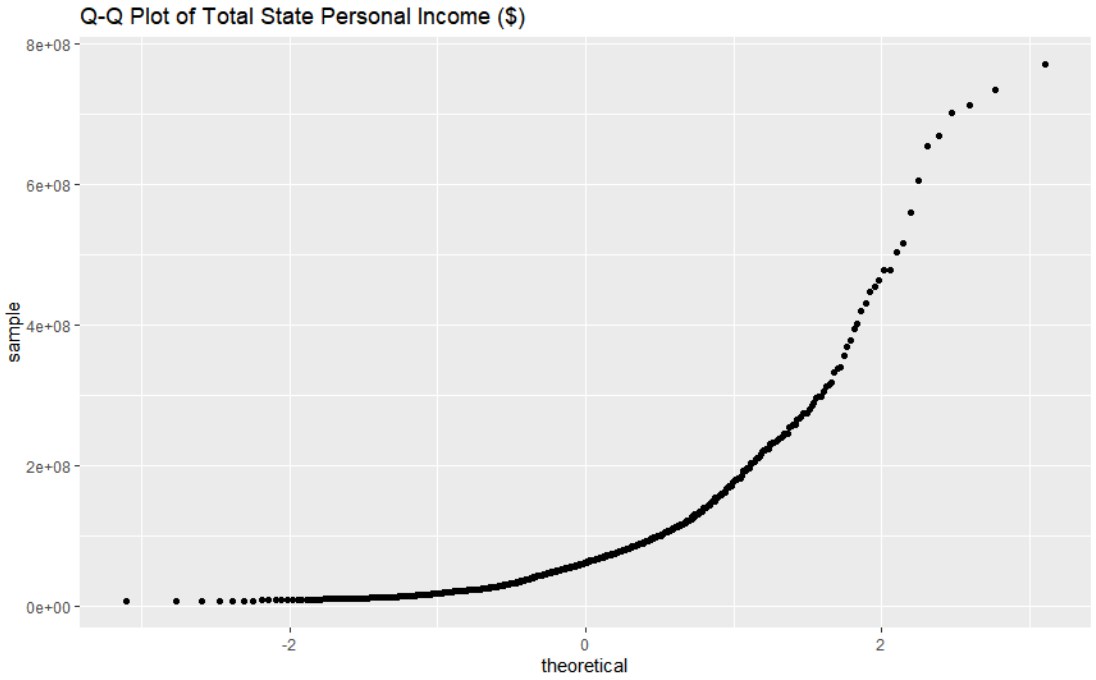
```
ggplot(Cigarette, aes(x = income, y = packpc)) +
geom_point(color = "darkorange1") +
  scale_x_log10() +
  geom_smooth(method = lm, color = "firebrick")
+
  xlab("Total State Personal Income (cents)") +
  ylab("Yearly Number of Packs Per Capita") +
  ggtitle("Total State Personal Income, Logarithmic
Scale ($) vs. Number of Packs Per Capita Per Year")
```

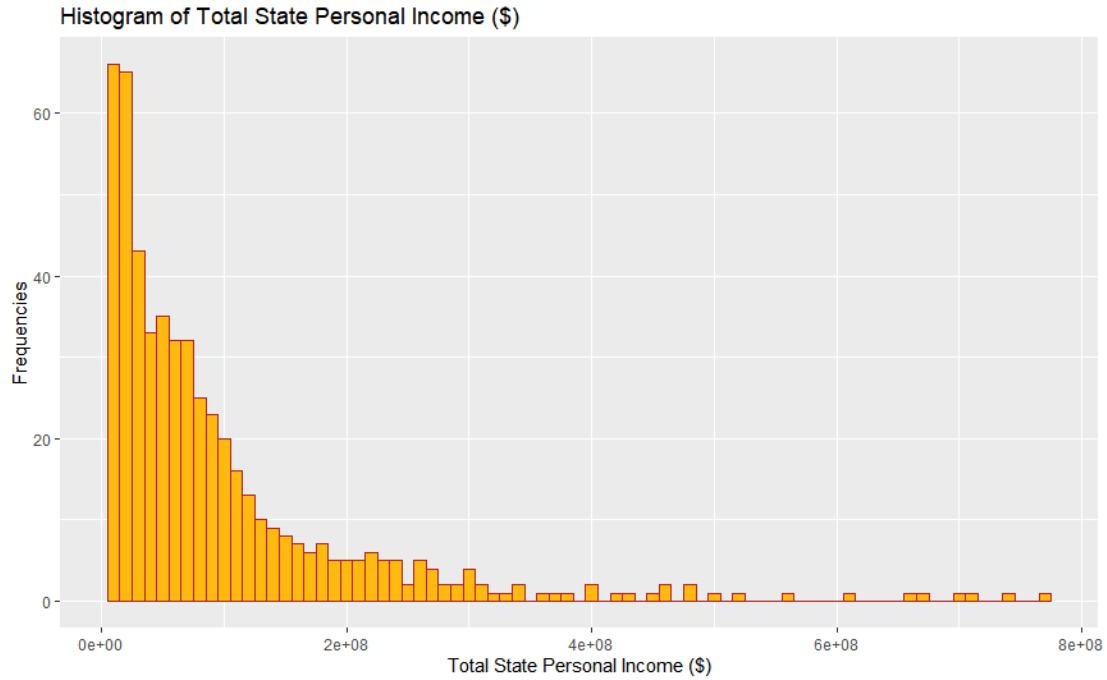


Correlation Between Packs Per Capita vs. Income (Logarithmic Scale)

- There appears to be a weak negative correlation between total state personal income and the yearly number of packs per capita.







Packs Per Capita vs. Income (Logarithmic Scale)

- The yearly average price per pack is normally distributed. However, the total state personal income is not normally distributed.

Correlation Between Packs Per Capita vs. Income (Logarithmic Scale)

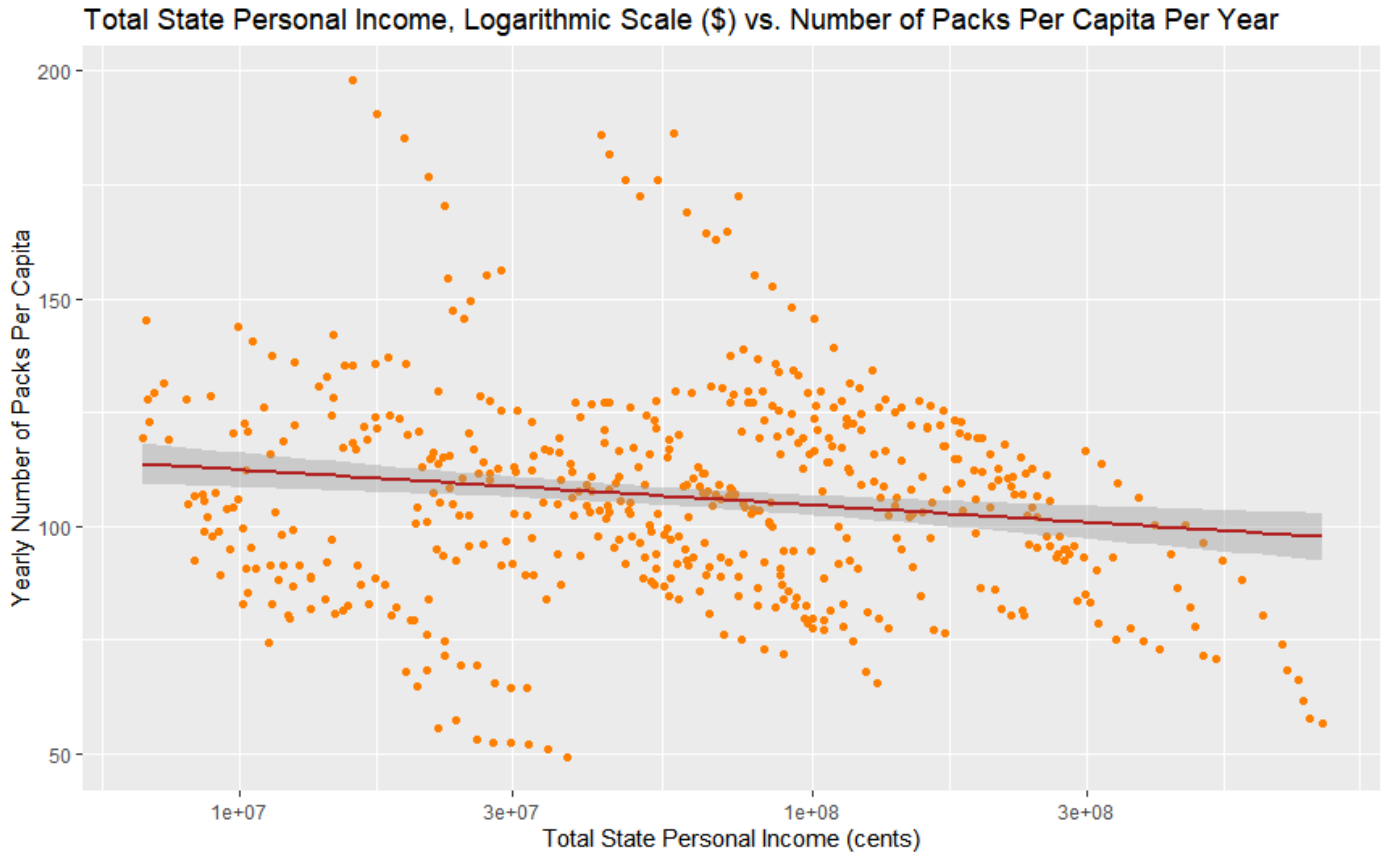
- Pearson's correlation coefficient, r:

-0.2553811

- Spearmans correlation coefficient, rho:

-0.1184955

- Both estimates of correlation coefficient show that there is a weak negative correlation between the total state personal income and the yearly number of packs per capita. The data shows that there is a slight connection between income and the prevalence in cigarette use.



Correlation Between Packs Per Capita vs. Income (Logarithmic Scale)

- Pearson's correlation coefficient, r:

-0.2553811

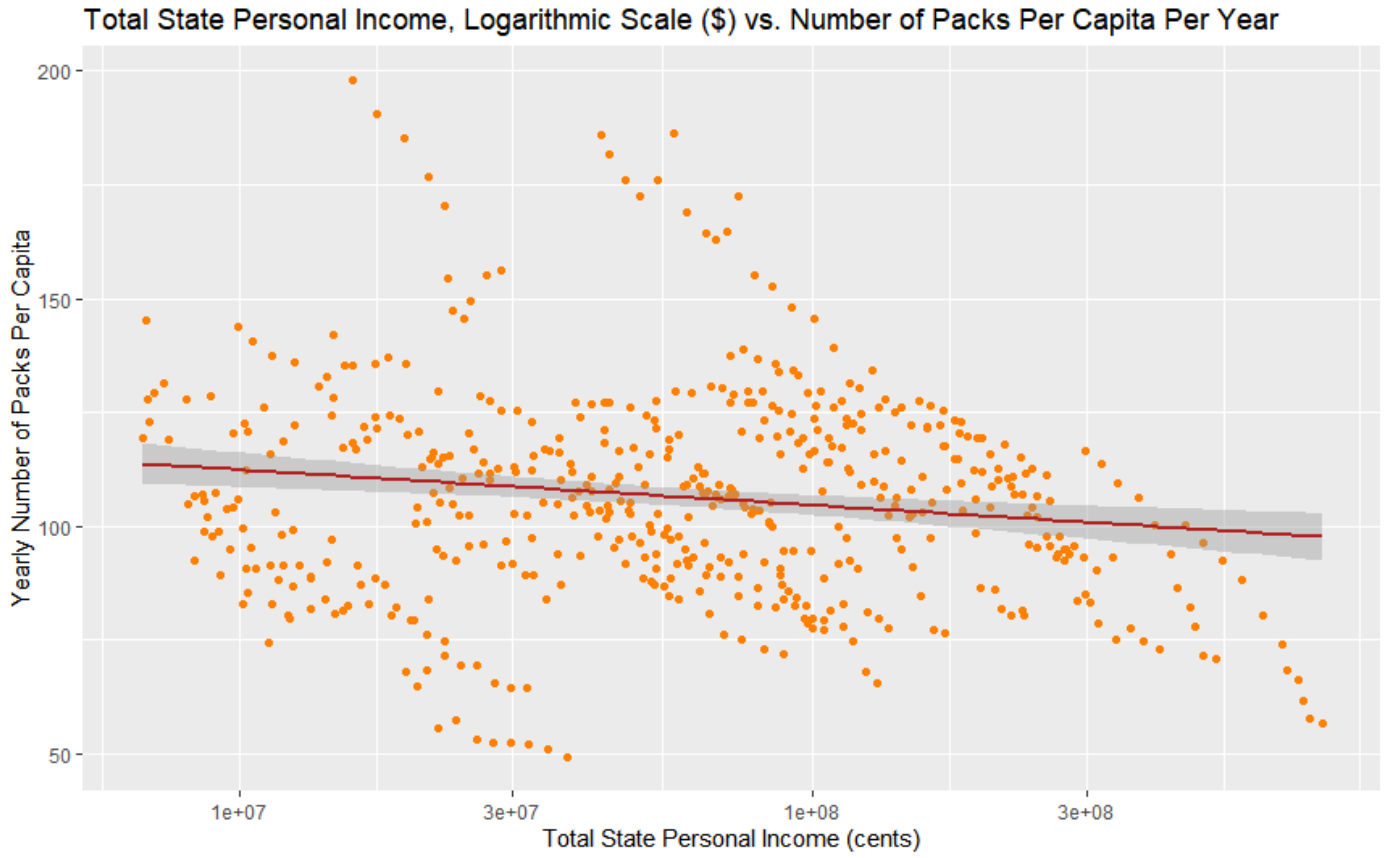
- Spearmans correlation coefficient, rho:

-0.1184955

Code to compute the correlation coefficient using both estimates:

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

```
cor.test(Cigarette$income, Cigarette$packpc, method = "spearman", use = "complete.obs")
```



Correlation Between Packs Per Capita vs. Income (Logarithmic Scale)

- The linear regression for the Between Packs Per Capita vs. Income is:

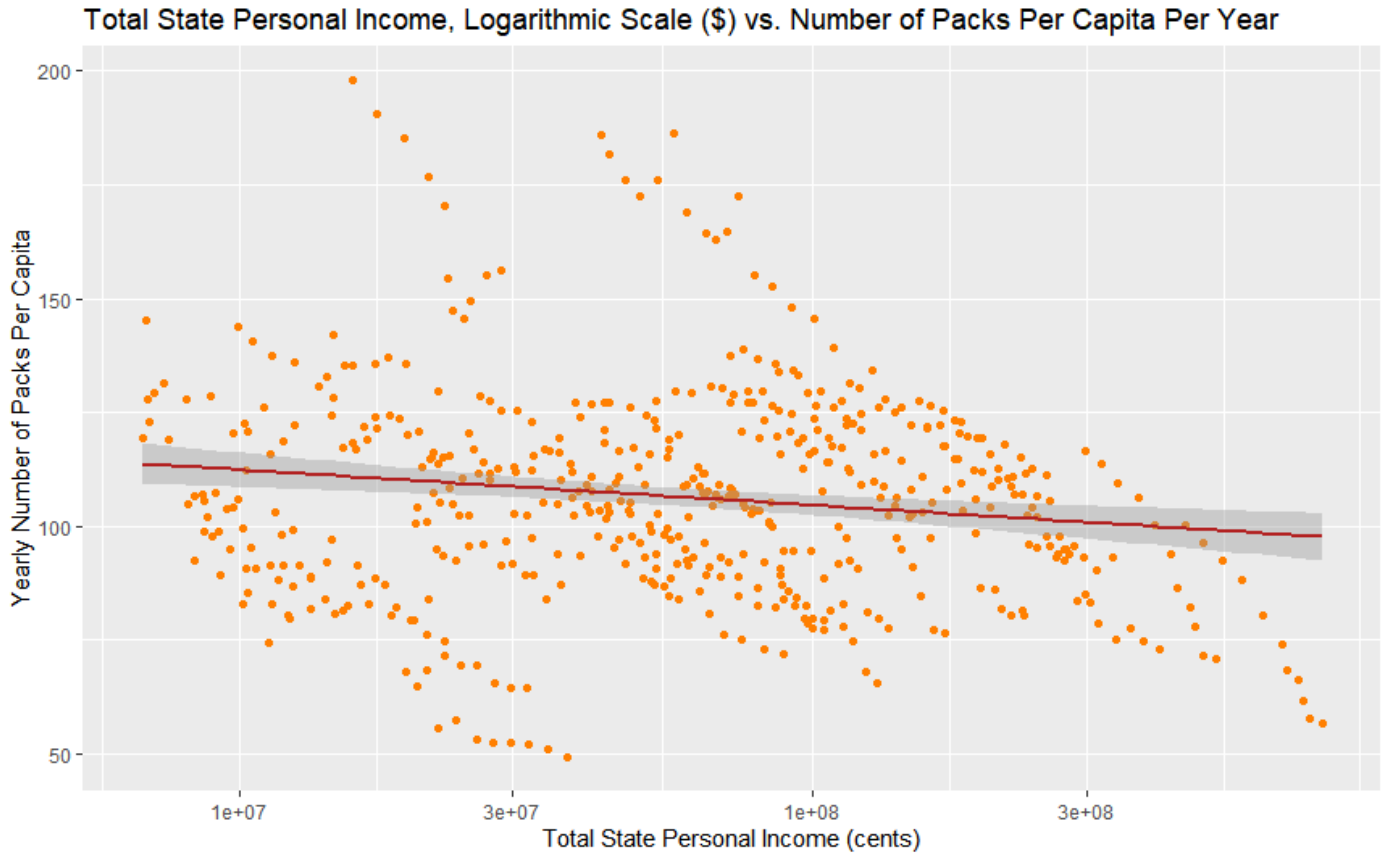
y = -5.047e-08 x + 1.115e+02

x = Total State Personal Income ($)

y = Yearly Number of Packs Per Capita

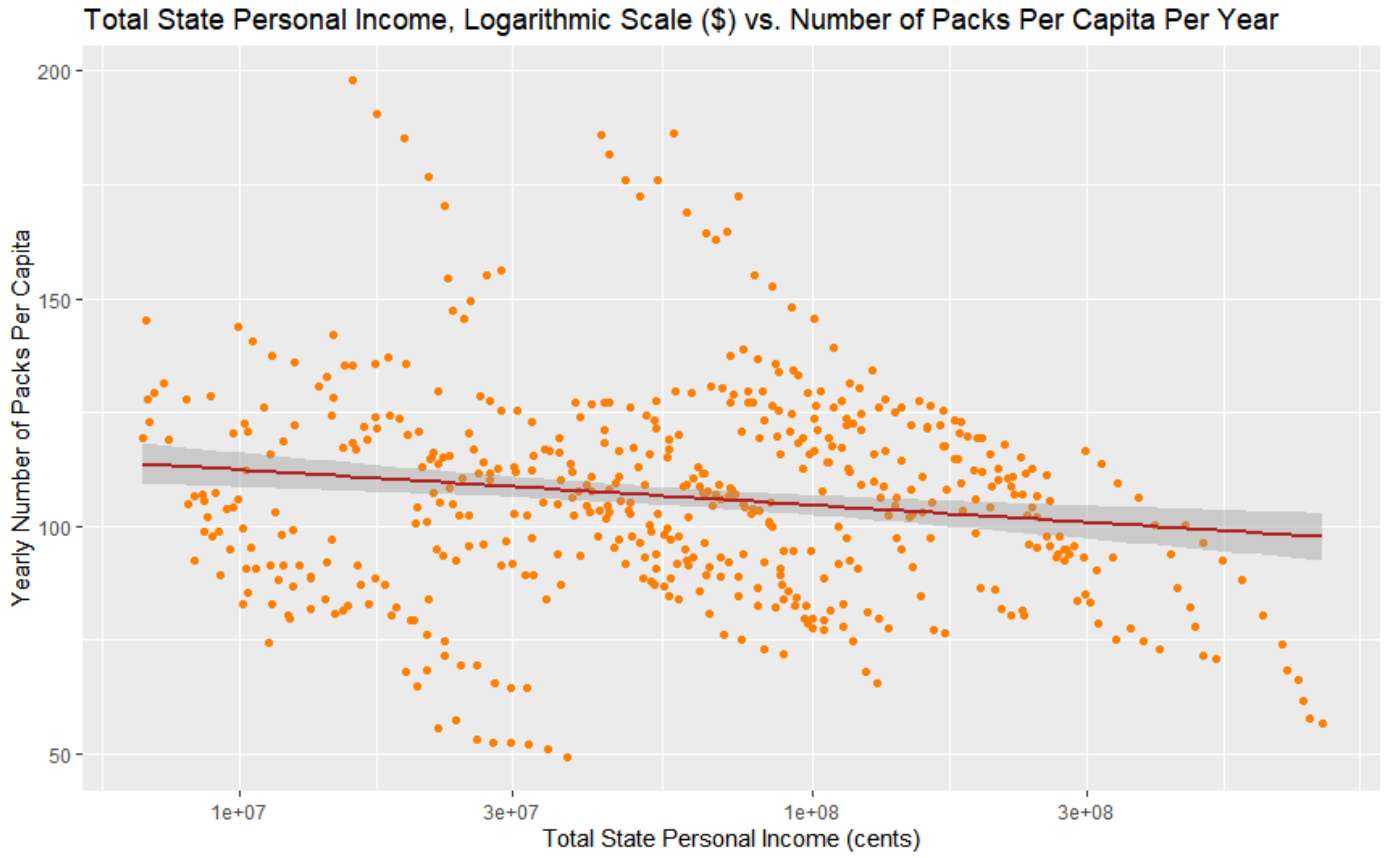
Code to compute this linear regression:

```
regression3 <- lm(packpc ~ income, Cigarette)
regression3
```



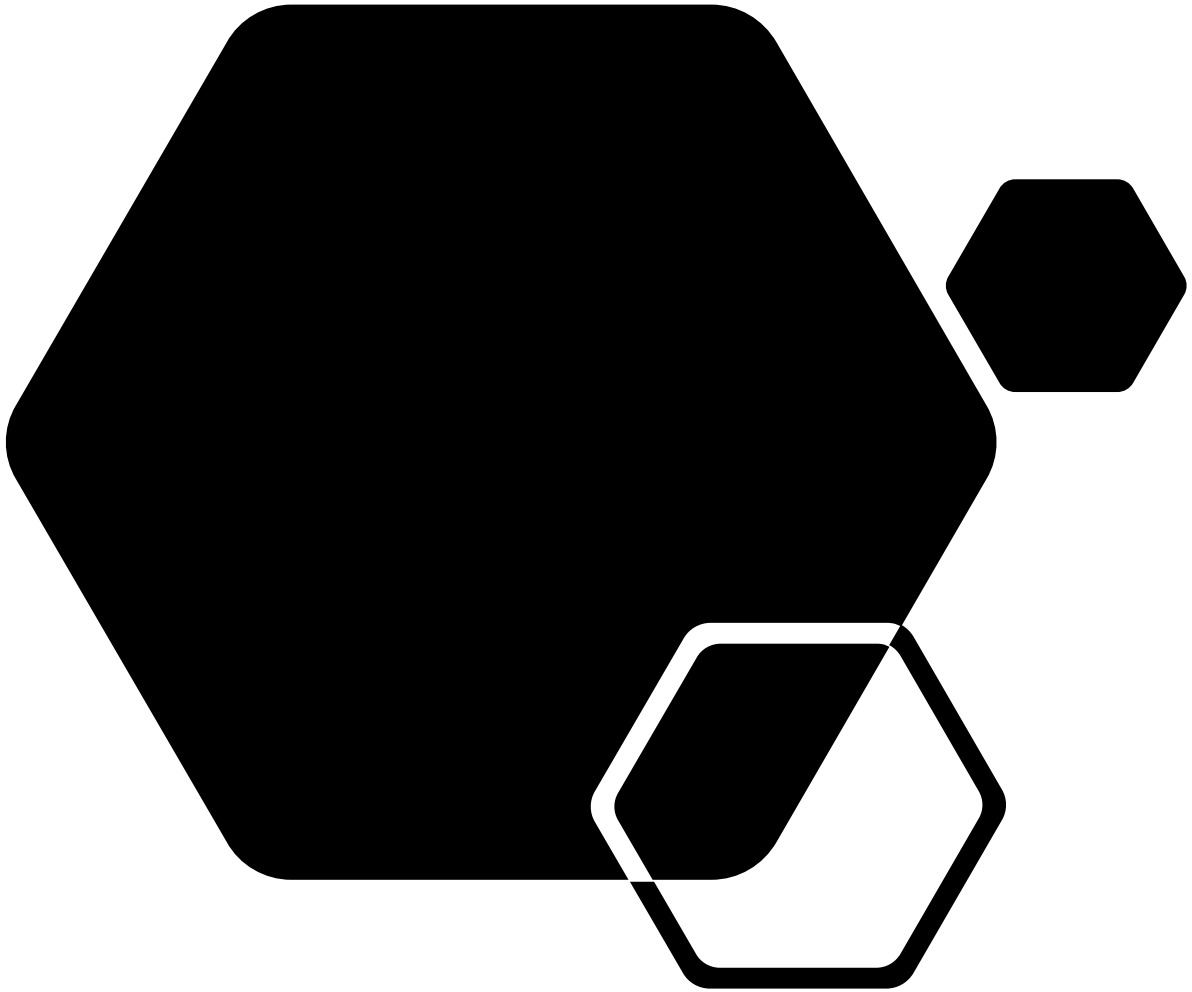
Correlation Between Packs Per Capita vs. Income (Logarithmic Scale)

- The p-value of the average price per pack is 2.63e-09, which is well below 0.05 and 0.001. Therefore, the regression is statistically significant.
- The adjusted r-squared value is 0.06344. This means that linear regression explains approximately 6.34% of the variability of the relationship between these two variables. Therefore, personal income level does not adequately explain the prevalence of cigarette use.
- The confidence interval narrows at the median personal income of the data and widens at the upper and lower personal income extremes. Therefore, the accuracy of the linear regression improves near the median average price per pack.



Part 7

Conclusion



Conclusion

- Between 1985 and 1995, the yearly number of packs per capita in the U.S. has been undergoing logistic decay.
- Between 1985 and 1995, Kentucky has the highest yearly number of packs per capita (173.90494), and Utah has the lowest (56.82223). Utah heavily taxes tobacco products.
- During 1989 – 1990, the U.S. experienced a sharp decline in cigarette consumption due taxes were levied on tobacco products. These taxes were then used to fund public health and environmental programs.
- As the average price per pack of cigarettes increases, the prevalence of cigarette consumption tends to decrease.
- Between 1985 and 1995, there is a dramatic decline in cigarette consumption.
- Lastly, there is a weak link between the total state personal income and the extent of cigarette consumption. Lower-income states tend to experience more cigarette usage.