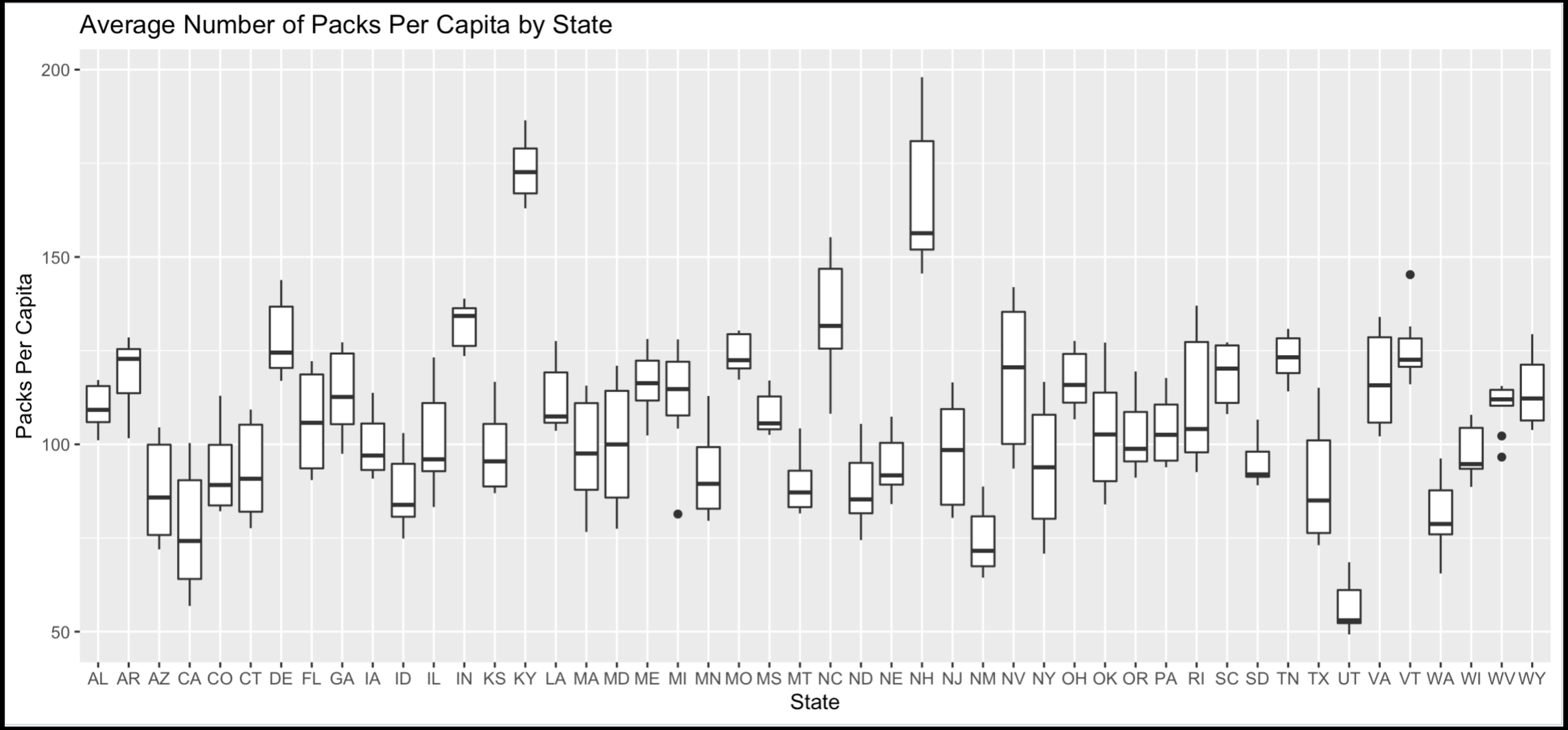
AN EXPLORATION OF THE CIGARETTE DATA SET

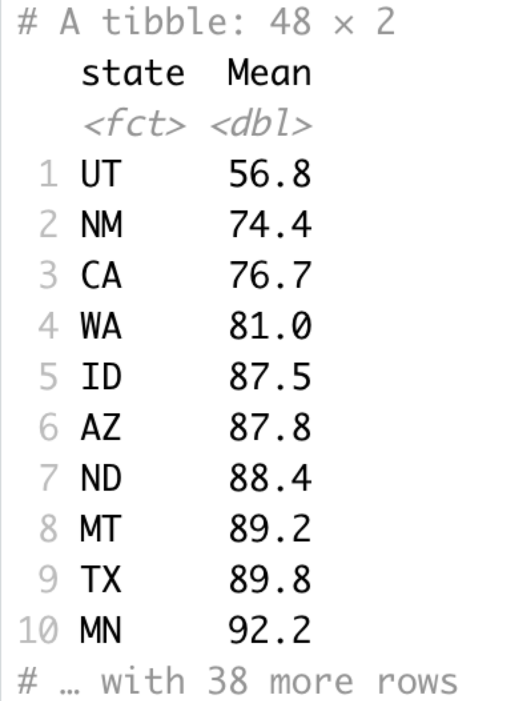
BOX PLOT OF THE AVERAGE NUMBER OF PACKS PER CAPITA BY STATE:



```
Cig.boxplot <- ggplot(Cigarette,aes(x = state, y = packpc)) + geom_boxplot() +
  xlab("State") + ylab("Packs Per Capita") +
  ggtitle("Average Number of Packs Per Capita by State")
```

REVIEW OF THE AVERAGE NUMBER OF PACKS PER CAPITA BY STATE

- The previous box plot had too many variables to sort through, so the information is organized to show the mean average number of packs per capita by state:

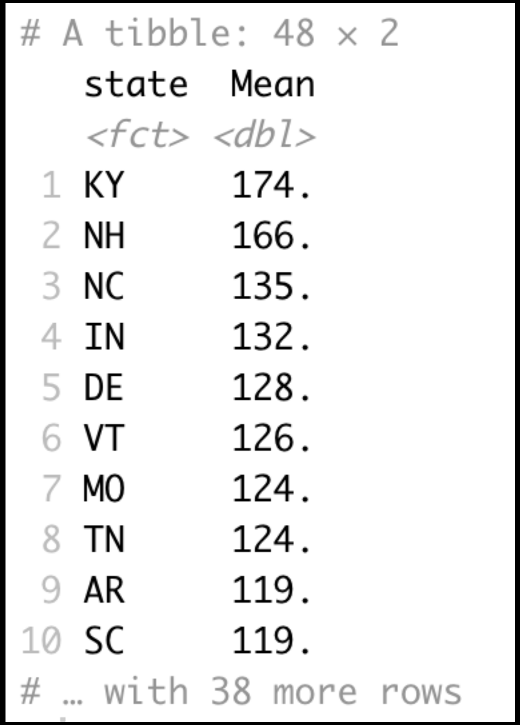


```
# A tibble: 48 × 2
   state  Mean
   <fct> <dbl>
 1 UT     56.8
 2 NM     74.4
 3 CA     76.7
 4 WA     81.0
 5 ID     87.5
 6 AZ     87.8
 7 ND     88.4
 8 MT     89.2
 9 TX     89.8
10 MN     92.2
# … with 38 more rows
```

< Lowest number of packs

Highest number of packs >

- Looking at the tibble on the left, the lowest number of packs per capita were in Utah. While the tibble on the right shows the highest number of packs per capita were in Kentucky.

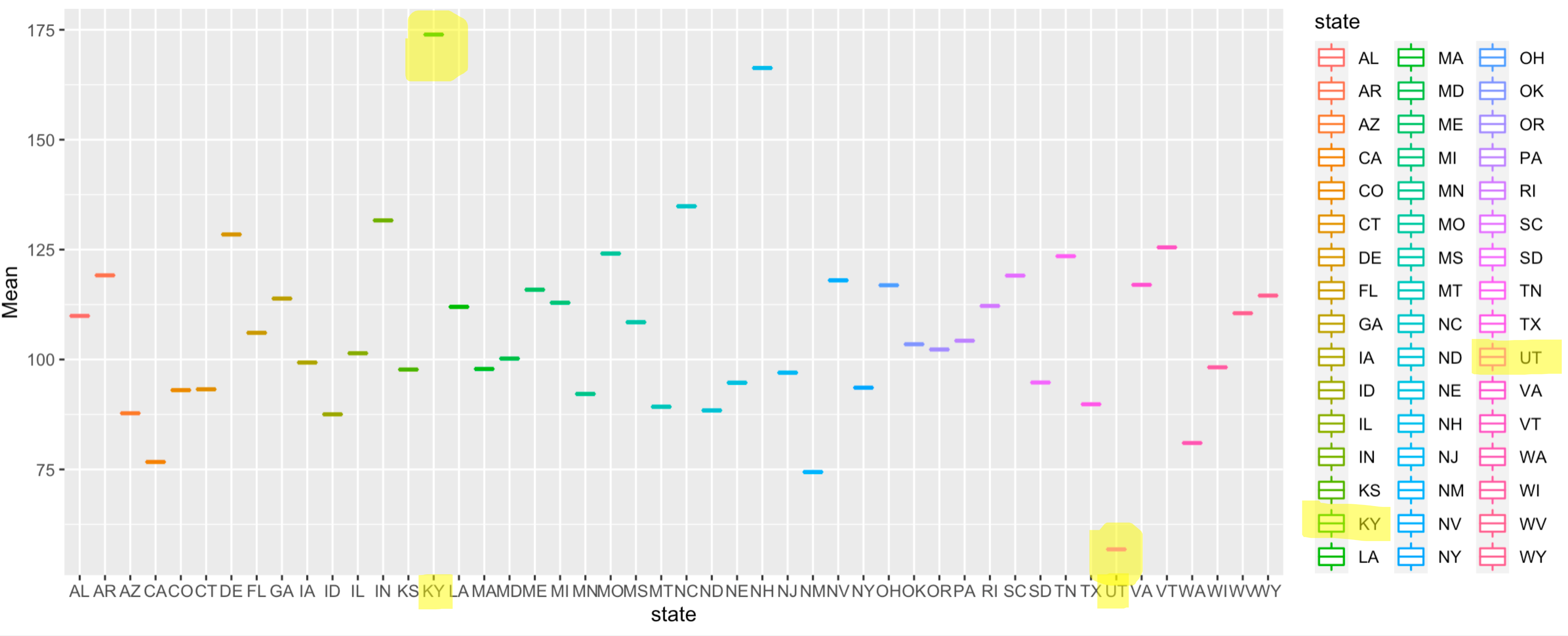


```
# A tibble: 48 × 2
   state  Mean
   <fct> <dbl>
 1 KY     174.
 2 NH     166.
 3 NC     135.
 4 IN     132.
 5 DE     128.
 6 VT     126.
 7 MO     124.
 8 TN     124.
 9 AR     119.
10 SC     119.
# … with 38 more rows
```

```
Cig.boxplotL <- Cigarette %>% group_by(state) %>%
summarise(Mean = mean(packpc)) %>% arrange(Mean)
```

```
Cig.boxplotH <- Cigarette %>% group_by(state) %>%
summarise(Mean = mean(packpc)) %>% arrange(desc(Mean))
```

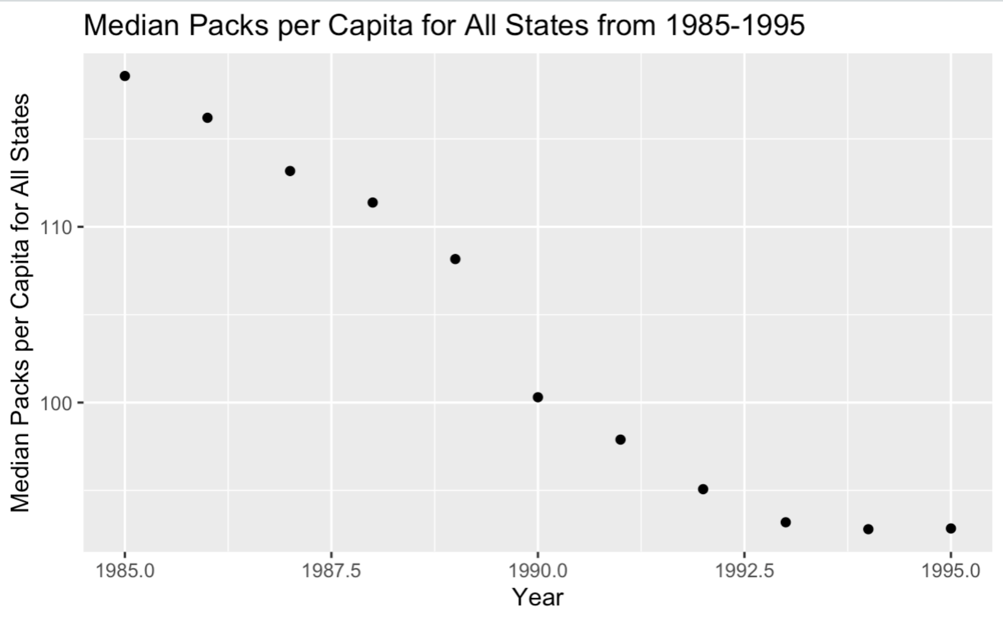
BOX PLOT OF THE MEAN OF PACKS PER CAPITA BY STATE:

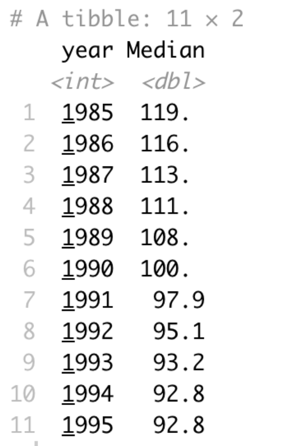


```
ggplot(Cig.boxplotL, aes(x = state, y = Mean, color = state)) + geom_boxplot()
```

MEDIAN OVER ALL THE STATES OF THE NUMBER OF PACKS PER CAPITA FOR EACH YEAR

- From 1985-1995 there is a steady decline of packs per capita each year. Note, starting in 1994, the packs per capita have about leveled off and continue with the same rate into 1995.



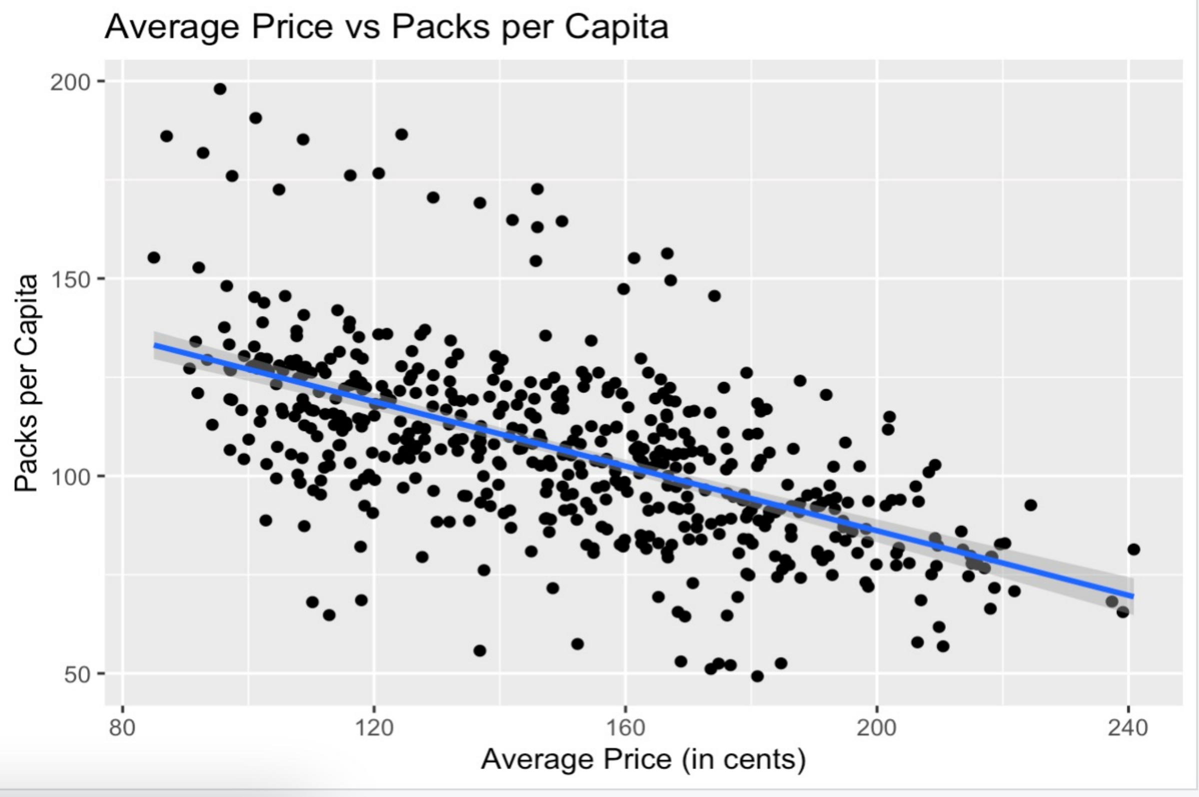


```
# A tibble: 11 × 2
    year Median
   <int>  <dbl>
 1  1985  119. 
 2  1986  116. 
 3  1987  113. 
 4  1988  111. 
 5  1989  108. 
 6  1990  100. 
 7  1991   97.9
 8  1992   95.1
 9  1993   93.2
10  1994   92.8
11  1995   92.8
```

```
CigMedian <- Cigarette %>% group_by(year) %>% summarise(Median = 
median(packpc))

CigMedYear <- ggplot(CigMedian, aes(x = year, y = Median)) + geom_point() +
  xlab("Year") +
  ylab("Median Packs per Capita for All States") +
  ggtitle("Median Packs per Capita for All States from 1985-1995")
```

SCATTER PLOT OF PRICE PER PACK VS. NUMBER OF PACKS PER CAPITA

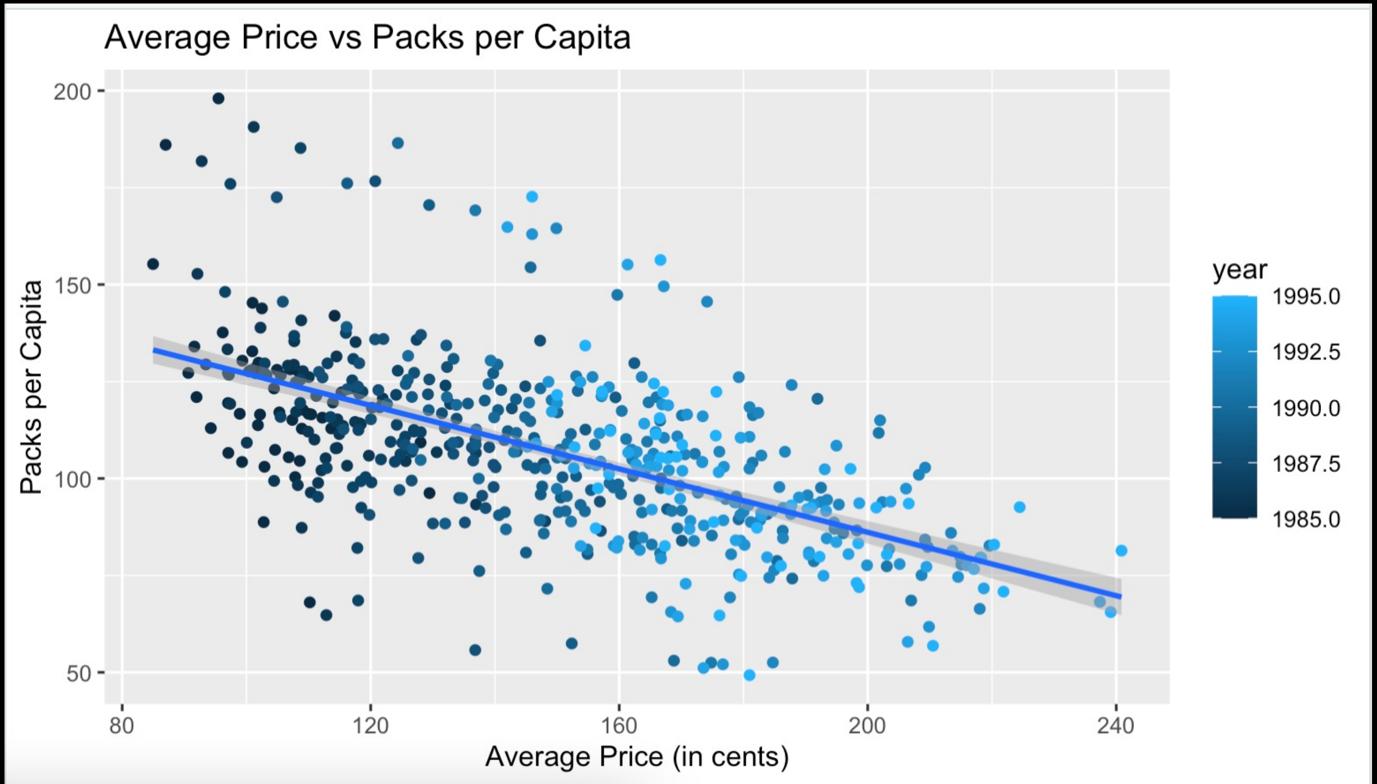
- The average price and the per capita are negatively correlated; this is expected as one would assume that as the price increases over the years, the packs bought would decrease.
- cor.test(Cigarette$avgprs, Cigarette$packpc, method = "pearson", use = "complete.obs")



```
CigScatter <- ggplot(Cigarette, aes(x = avgprs, y = packpc)) + geom_point() +
  geom_smooth(method = lm) +
  xlab("Average Price (in cents)") + ylab("Packs per Capita") +
  ggtitle("Average Price vs Packs per Capita")
```

SCATTER PLOT: EMPHASISING YEAR

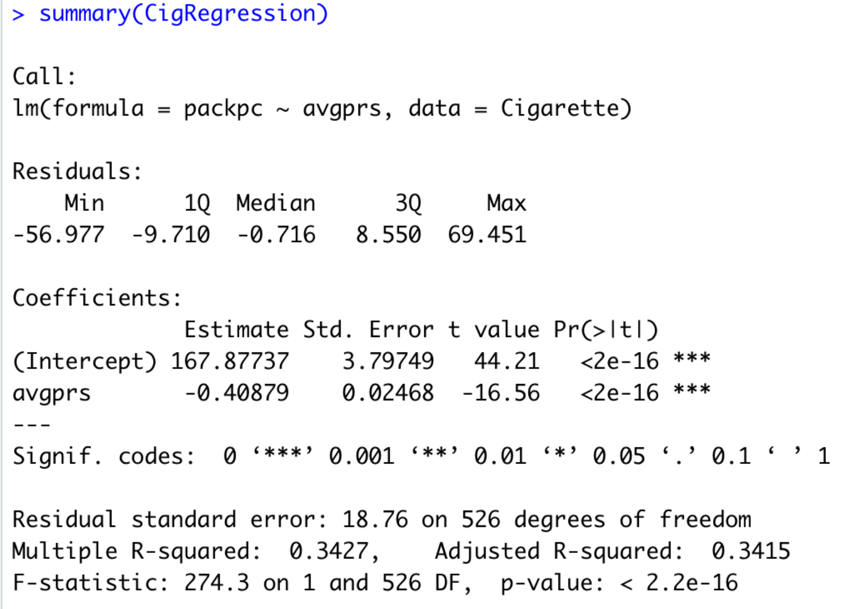
- The relationship between the two variables do change over time. Starting in 1985, when Cigarettes were less expensive, there were more packs per capita. Whereas later in the data set the average price has increased and the packs per capita has decreased.



```
CigScatterYear <- ggplot(Cigarette, aes(x = avgprs, y = packpc, color = year)) +
geom_point() +
  geom_smooth(method = lm) +
  xlab("Average Price (in cents)") + ylab("Packs per Capita") +
  ggtitle("Average Price vs Packs per Capita")
```

LINEAR REGRESSION OF PACKS PER CAPITA ~ AVERAGE PRICE

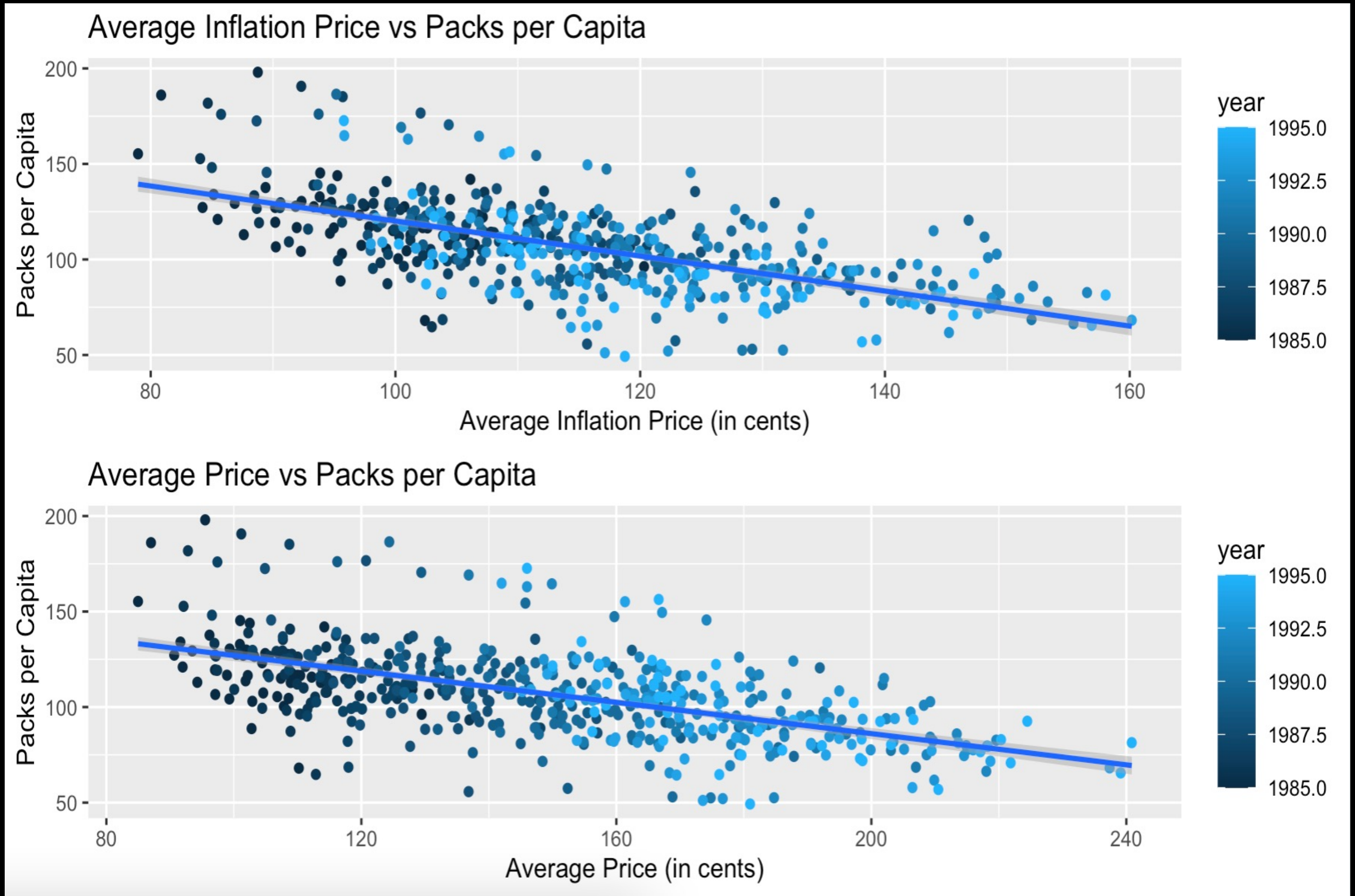
- 34% of the variability
- Packs per capita are going to decrease by -0.41 for every average one unit price increase.
- Average price per pack accounts for 34% of everything that influences the packs per capita.
- The p-value is <0.05 so the overall model is significant. Price per pack is a significant predictor of the packs per capita. The higher the price is, the lower the number of packs of cigarettes are sold.



```
> summary(CigRegression)

Call:
lm(formula = packpc ~ avgprs, data = Cigarette)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.977  -9.710  -0.716   8.550  69.451 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 167.87737    3.79749   44.21   <2e-16 ***
avgprs       -0.40879    0.02468  -16.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.76 on 526 degrees of freedom
Multiple R-squared:  0.3427,	Adjusted R-squared:  0.3415 
F-statistic: 274.3 on 1 and 526 DF,  p-value: < 2.2e-16
```

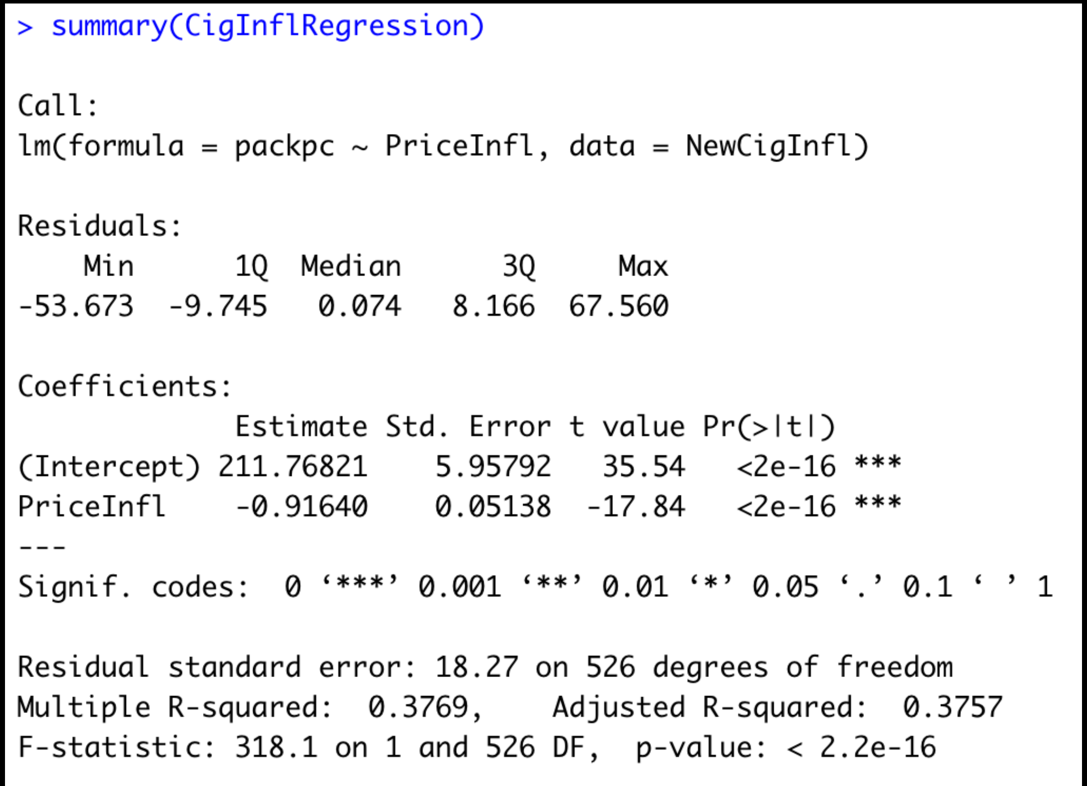
CigRegression <- lm(packpc ~ avgprs, Cigarette)
summary(CigRegression)

SCATTER PLOT: PRICE ADJUSTED FOR INFLATION



- NewCigInfl <- Cigarette %>% mutate(PriceInfl = avgprs/cpi)
- CigInflScatter <- ggplot(NewCigInfl, aes(x = PriceInfl, y = packpc)) + geom_point() + geom_smooth(method = lm) + xlab("Average Inflation Price (in cents)") + ylab("Packs per Capita") + ggtitle("Average Inflation Price vs Packs per Capita")
- grid.arrange(CigInflScatterYear, CigScatterYear, ncol = 1)

LINEAR REGRESSION: PRICE ADJUSTED FOR INFLATION

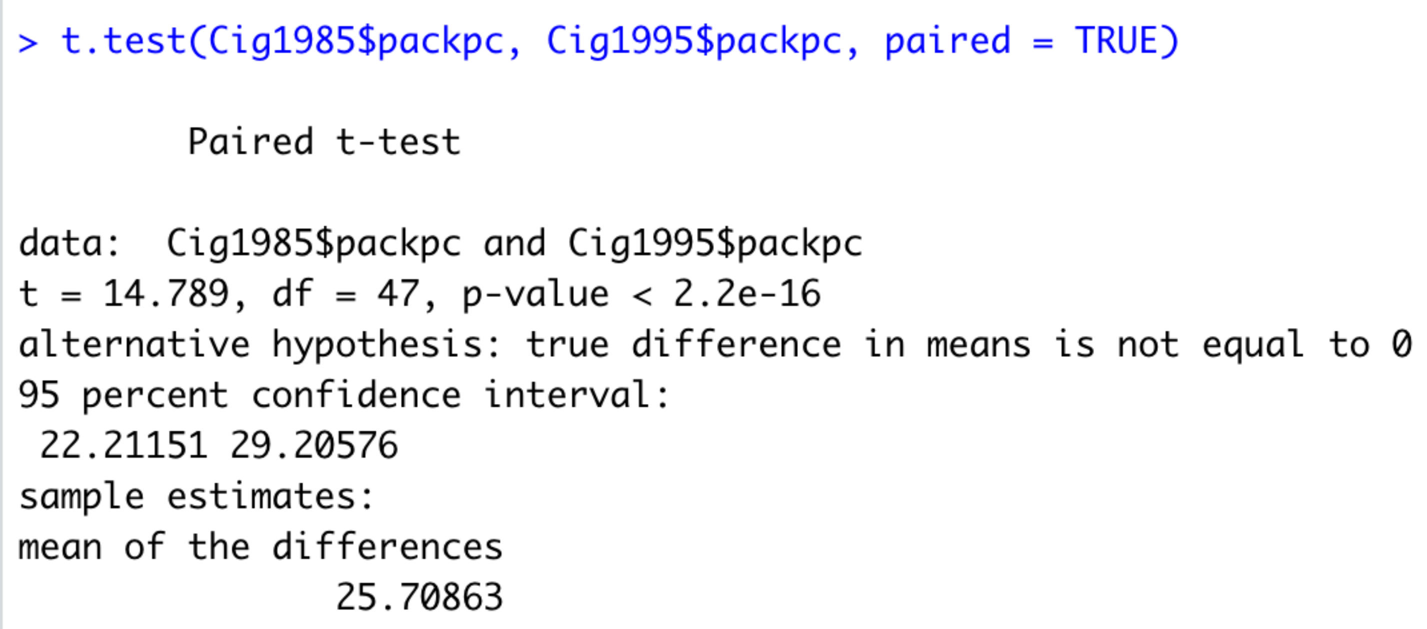


```
> summary(CigInflRegression)

Call:
lm(formula = packpc ~ PriceInfl, data = NewCigInfl)

Residuals:
    Min      1Q  Median      3Q     Max 
-53.673  -9.745   0.074   8.166  67.560 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 211.76821    5.95792   35.54   <2e-16 ***
PriceInfl    -0.91640    0.05138  -17.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.27 on 526 degrees of freedom
Multiple R-squared:  0.3769,	Adjusted R-squared:  0.3757 
F-statistic: 318.1 on 1 and 526 DF,  p-value: < 2.2e-16
```

CigInflRegression <- lm(packpc ~ PriceInfl, NewCigInfl)
summary(CigInflRegression)

- 38% of the variability
- The p-value <0.05, thus the price still shows a significant impact on the packs per capita. When comparing the two graphs, both have a negative correlation and both can be said that the higher the price is, the lesser packs per capita.
- Note: inflation seems to have an effect on where the years lie on the graph of the Cigarette data set

PAIRED *T*-TEST: DIFFERENCE BETWEEN 1985 & 1995



```
> t.test(Cig1985$packpc, Cig1995$packpc, paired = TRUE)

        Paired t-test

data:  Cig1985$packpc and Cig1995$packpc
t = 14.789, df = 47, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 22.21151 29.20576
sample estimates:
mean of the differences
               25.70863
```

- p < .05; there is a significant difference between packs per capita in 1985 and packs per capita in 1995

```
Cig1985 <- Cigarette %>% filter(year == "1985")

Cig1995 <- Cigarette %>% filter(year == "1995")

t.test(Cig1985$packpc, Cig1995$packpc, paired = TRUE)
```

CURIOSITIES FROM THE CIGARETTES DATA SET

- When looking at the median packs per capita, what event happened to cause such a rapid decline from 1989 and onward?
  - Public health programs became a priority
  - A focus on protecting non-users from second-hand smoke was another factor that helped to usher in the decline of cigarette use in 1993
  - Smoking restrictions in public places, such as in restaurants and at work, became more prevalent