Project 3

Megan Oh, Aiden Mayhood, Tiantian Zhao, Han Zheng

2022-11-15

Project 3

I. Panel Data Model

```
# download panel data set
data(CigarettesSW)
state <- CigarettesSW$state
year <- CigarettesSW$year
cpi <- CigarettesSW$cpi
population <- CigarettesSW$population
packs <- CigarettesSW$packs
income <- CigarettesSW$income
tax <- CigarettesSW$tax
price <- CigarettesSW$price
taxs <- CigarettesSW$taxs
data <- data.frame(state, year, cpi, population, packs, income,
    tax, price, taxs)
```

(1) This panel data set is from AER, and gives data on cigarette consumption from the 48 US continental states in 1985 and 1995. N=48 and T=2, indicating that our data frame is short and wide Also, our data set is considered balanced, as there are the same number of observations for each state across time. There are 7 different variables used to measure this data. The 6 explanatory variables are "cpi" which is the consumer price index for cigarettes, "population" which is that state's population, "income" which is the total personal, nominal income, "tax" which is the average state, federal, and local taxes for a fiscal year, "price" which is the average price of a pack of cigarettes, and "taxs" which is the average excise and sales tax for a fiscal year. The response variable for this data set is "packs" which is the number of packs of cigarettes consumed in that state per capita.

With our model we are trying to model how different factors, being our predictor variables, affect cigarette consumption in different states and in different years.

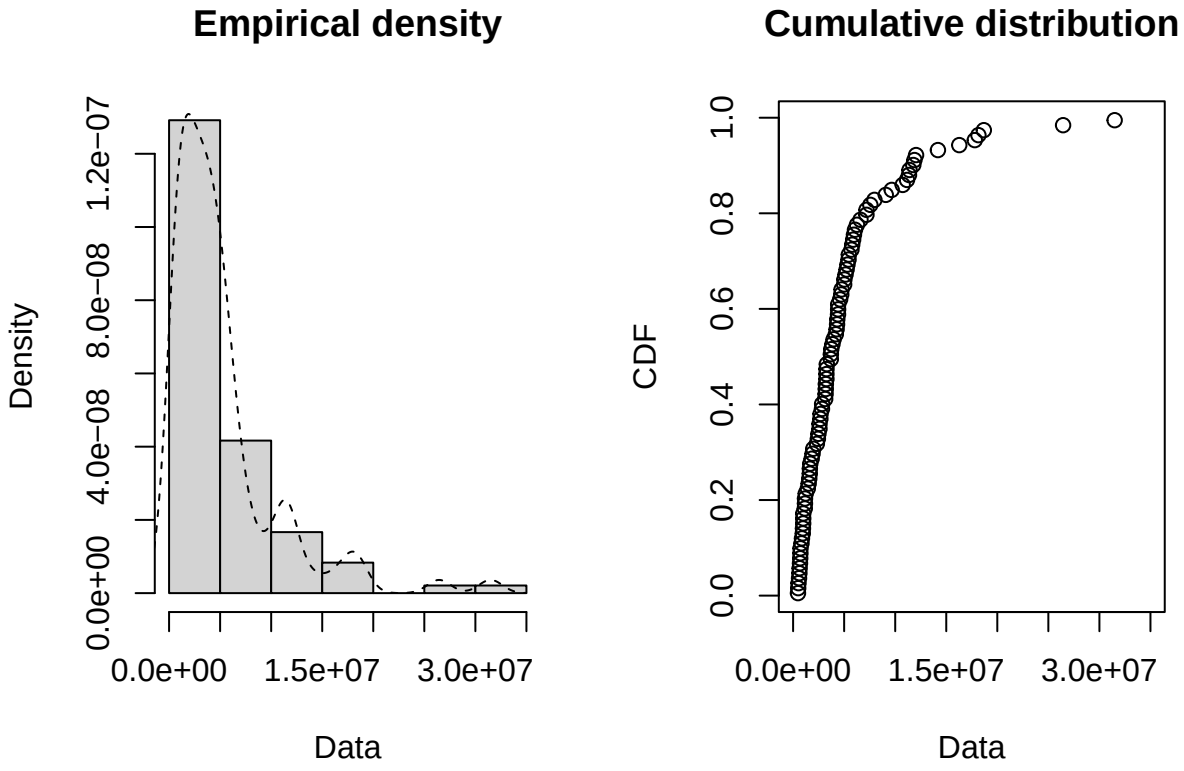
(2) Provide a descriptive analysis of all of the variables, including histograms, fitted distributions, correlation plots, box plots, scatterplots and statistical summaries.

The consumer price index for all of the states in the first time period, 1985, was 1.076, and the consumer price index for all of the states in the second period, 1995, was 1.524.

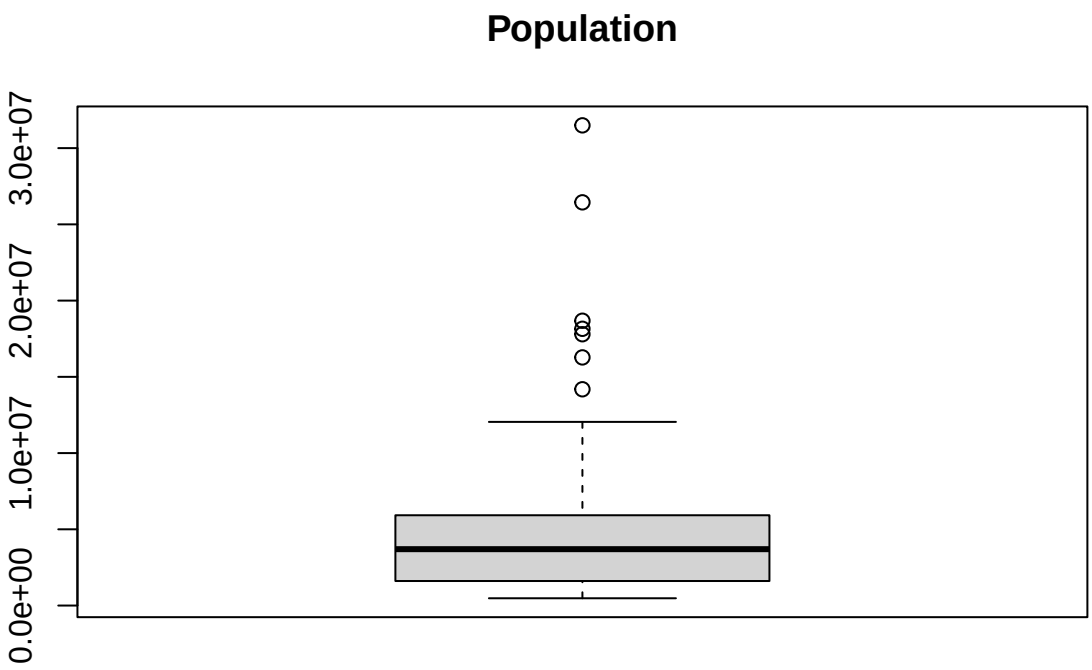
```
# provide five-number for each variable
summary(data)
```

```
##      state      year          cpi          population           packs
##  AL     : 2   1985:48   Min.   :1.076   Min.   :  478447   Min.   : 49.27
##  AR     : 2   1995:48   1st Qu.:1.076   1st Qu.: 1622606   1st Qu.: 92.45
##  AZ     : 2             Median :1.300   Median : 3697472   Median :110.16
##  CA     : 2             Mean   :1.300   Mean   : 5168866   Mean   :109.18
##  CO     : 2             3rd Qu.:1.524   3rd Qu.: 5901500   3rd Qu.:123.52
##  CT     : 2             Max.   :1.524   Max.   :31493524   Max.   :197.99
##  (Other):84
##      income                tax             price             taxs
##  Min.   :  6887097   Min.   :18.00   Min.   : 84.97   Min.   : 21.27
##  1st Qu.: 25520384   1st Qu.:31.00   1st Qu.:102.71   1st Qu.: 34.77
##  Median : 61661644   Median :37.00   Median :137.72   Median : 41.05
##  Mean   : 99878736   Mean   :42.68   Mean   :143.45   Mean   : 48.33
##  3rd Qu.:127313964   3rd Qu.:50.88   3rd Qu.:176.15   3rd Qu.: 59.48
##  Max.   :771470144   Max.   :99.00   Max.   :240.85   Max.   :112.63
##
```

```
# provide a histogram and box plot for the population
plotdist(data$population, histo = TRUE, demp = TRUE)
```

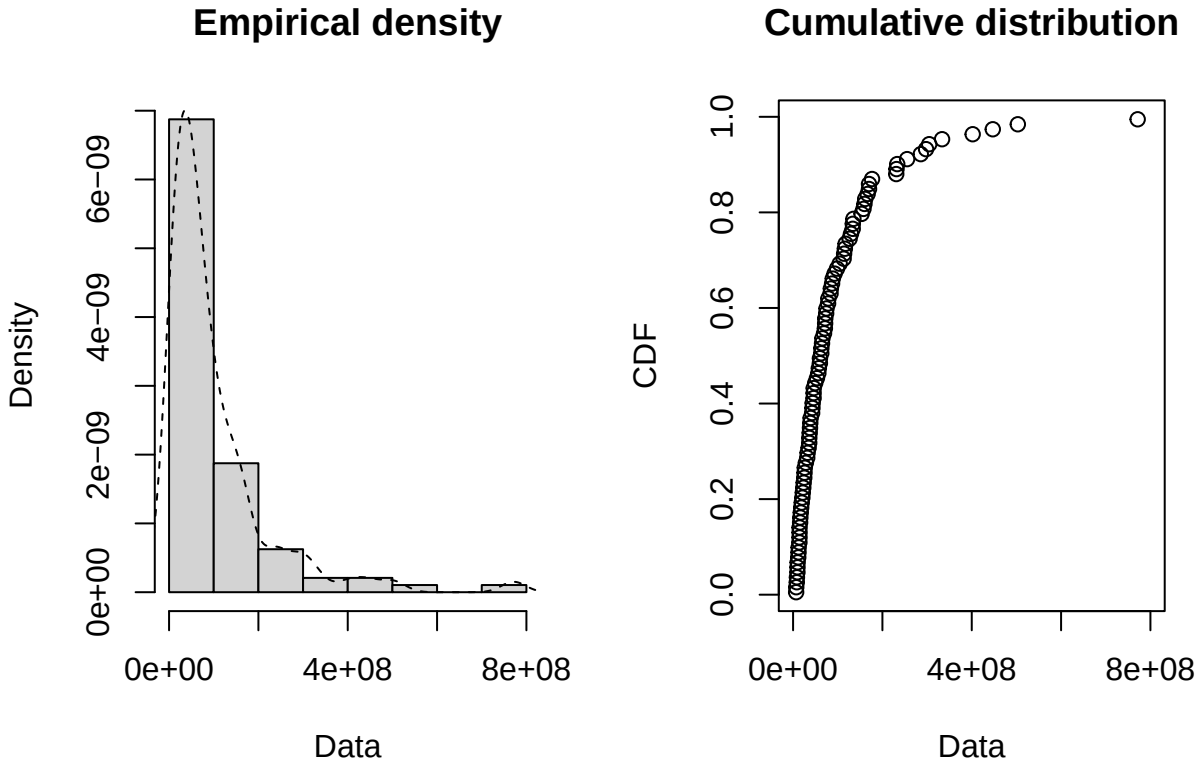



```
boxplot(data$population, main = "Population")
```

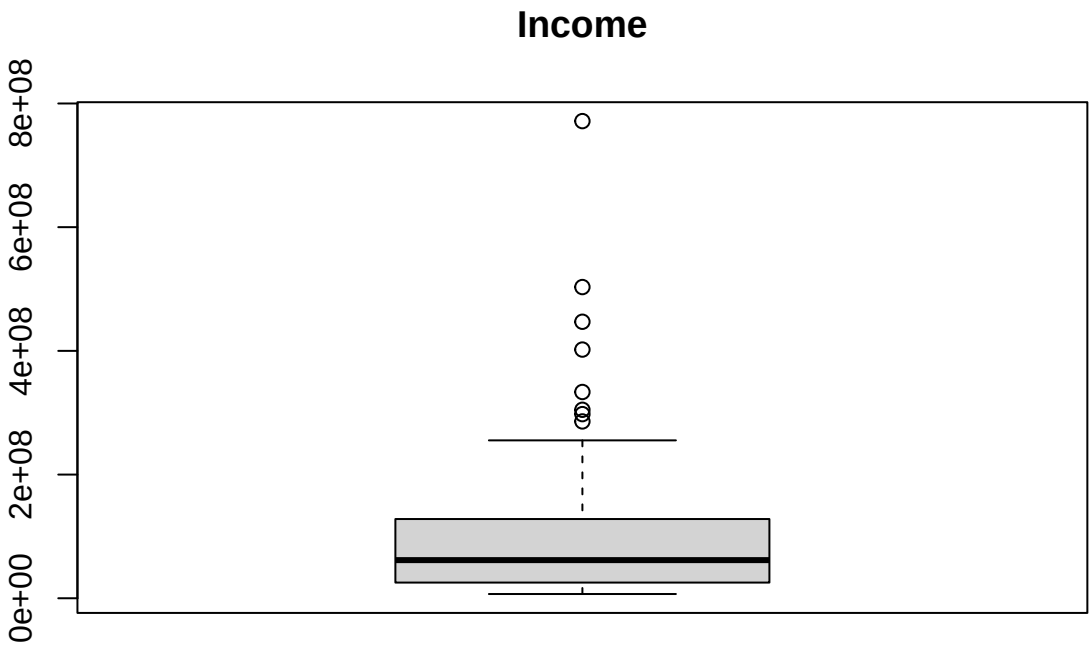


The histogram for population shows that the data is heavily skewed right, with most of the data being relatively low. Furthermore, from the box plot we can see that population has a few outliers on the higher side.

```
# provide a histogram and box plot for income
plotdist(data$income, histo = TRUE, demp = TRUE)
```

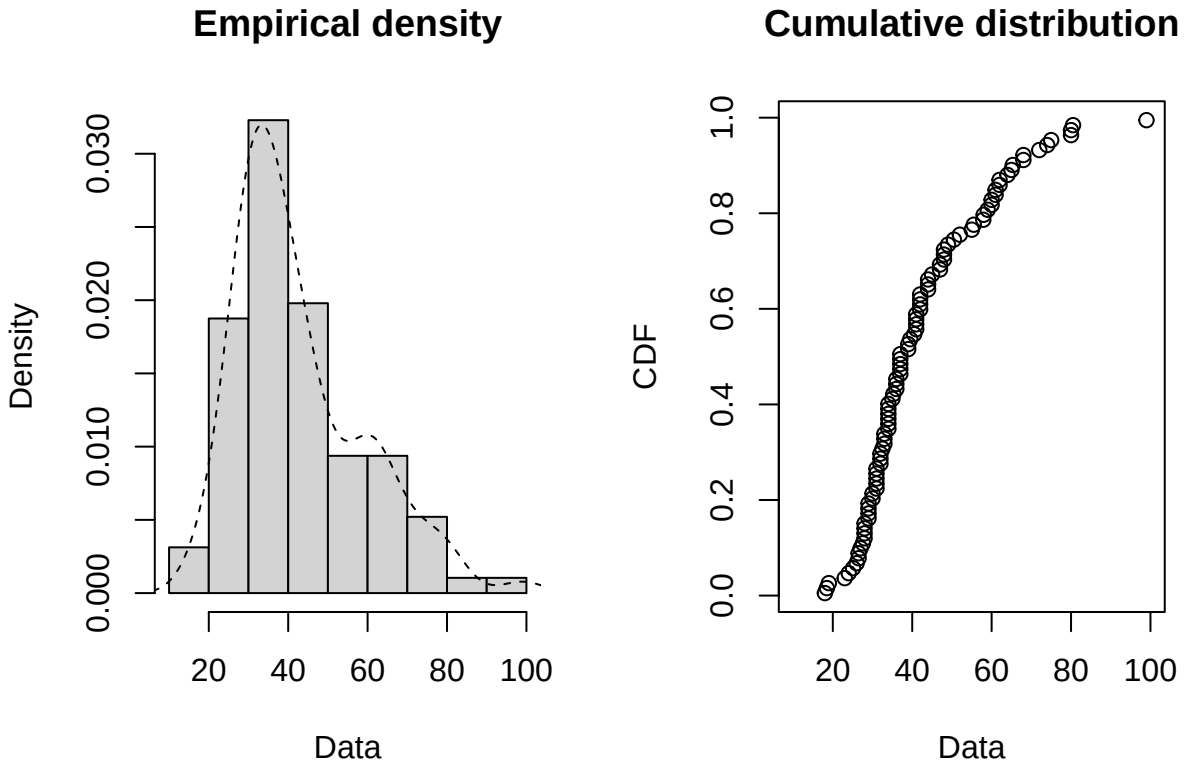


```
boxplot(data$income, main = "Income")
```

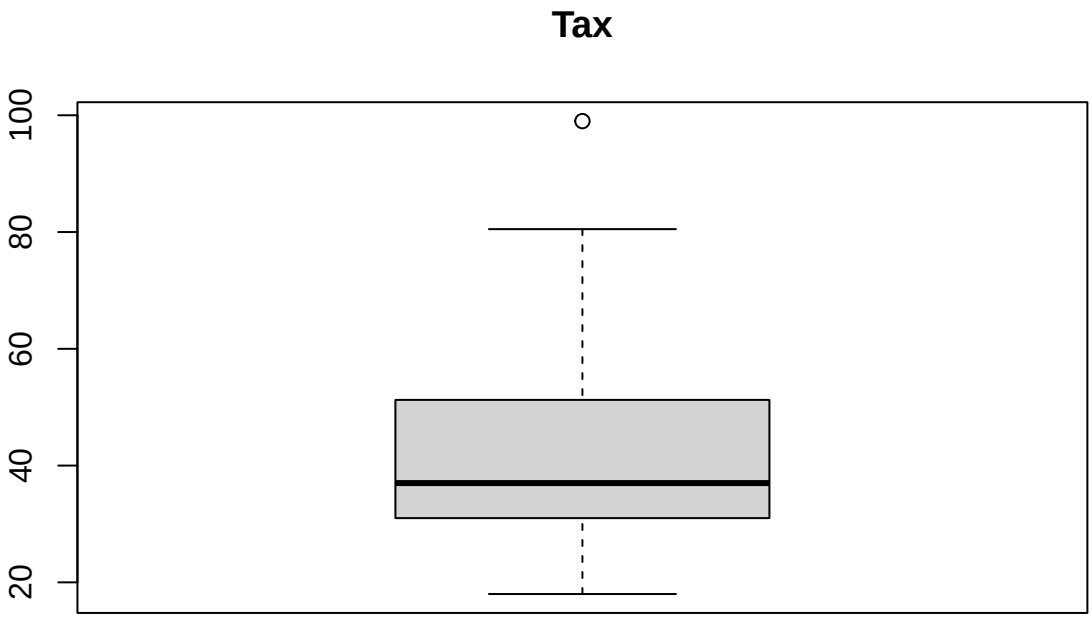


The histogram for population shows that it heavily skewed right. In addition, the box plot shows that income has a few outliers.

```
# provide a histogram and box plot for tax
plotdist(data$tax, histo = TRUE, demp = TRUE)
```

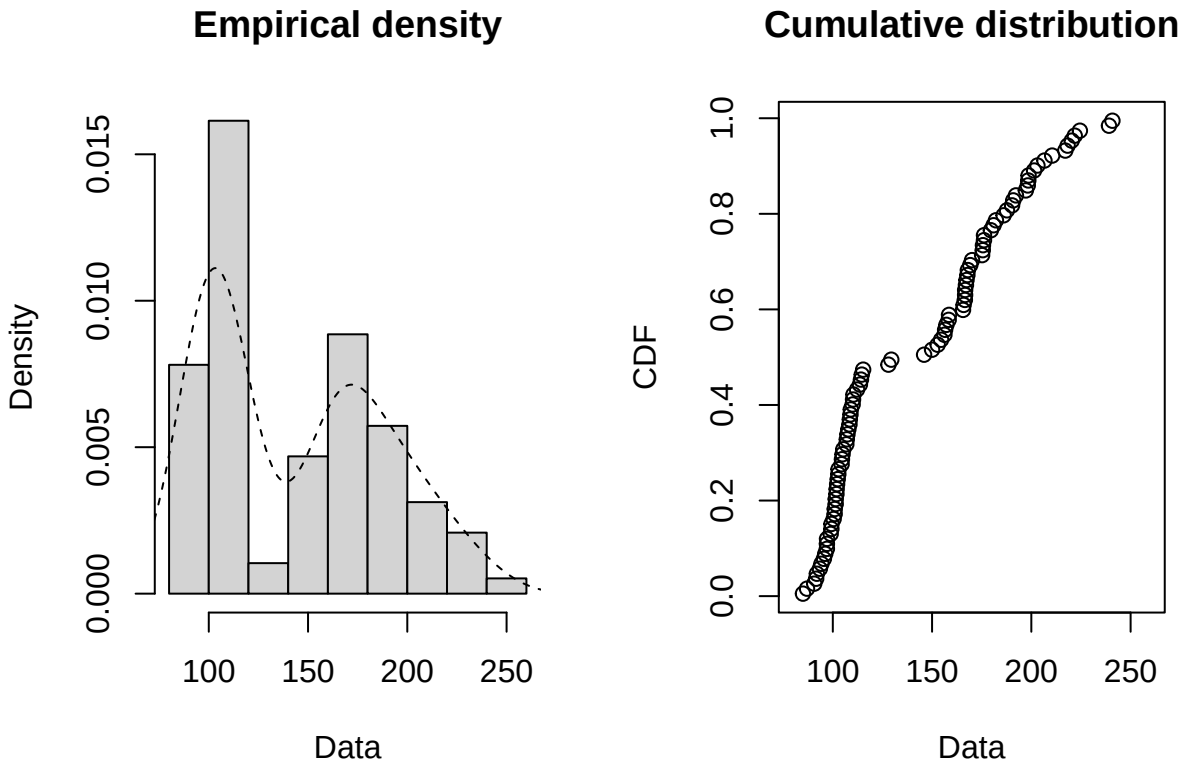


```
boxplot(data$tax, main = "Tax")
```

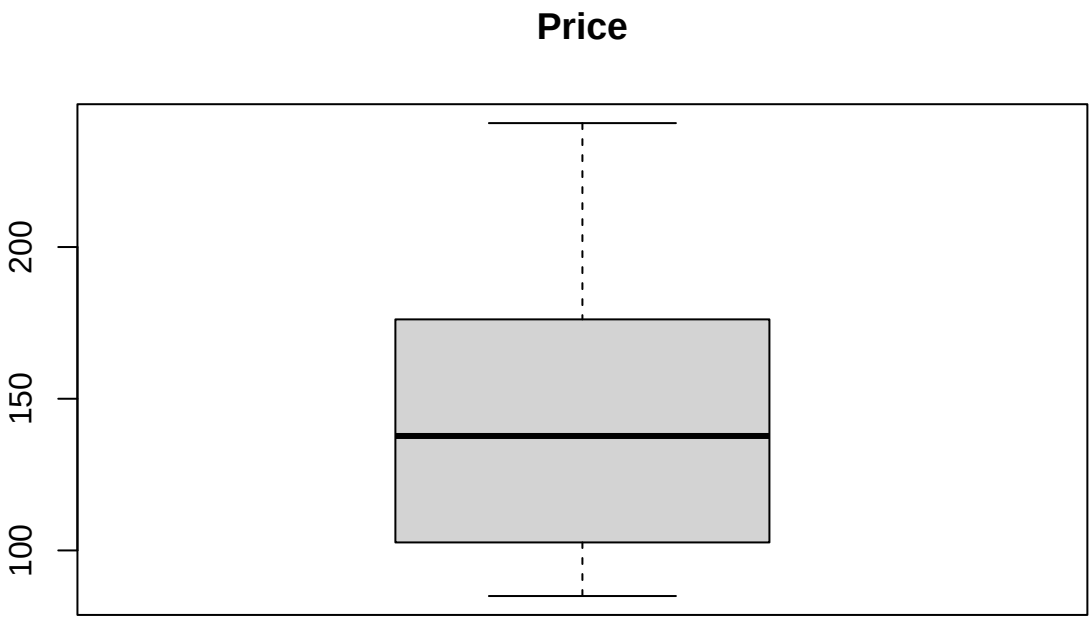


The histogram for tax shows that it is slightly skewed right, with most values being around 30-40. In addition, the box plot shows that tax has one outlier that is higher than the other tax values, and that tax ranges from approximately 20 to 80.

```
# provide a histogram and box plot for price
plotdist(data$price, histo = TRUE, demp = TRUE)
```

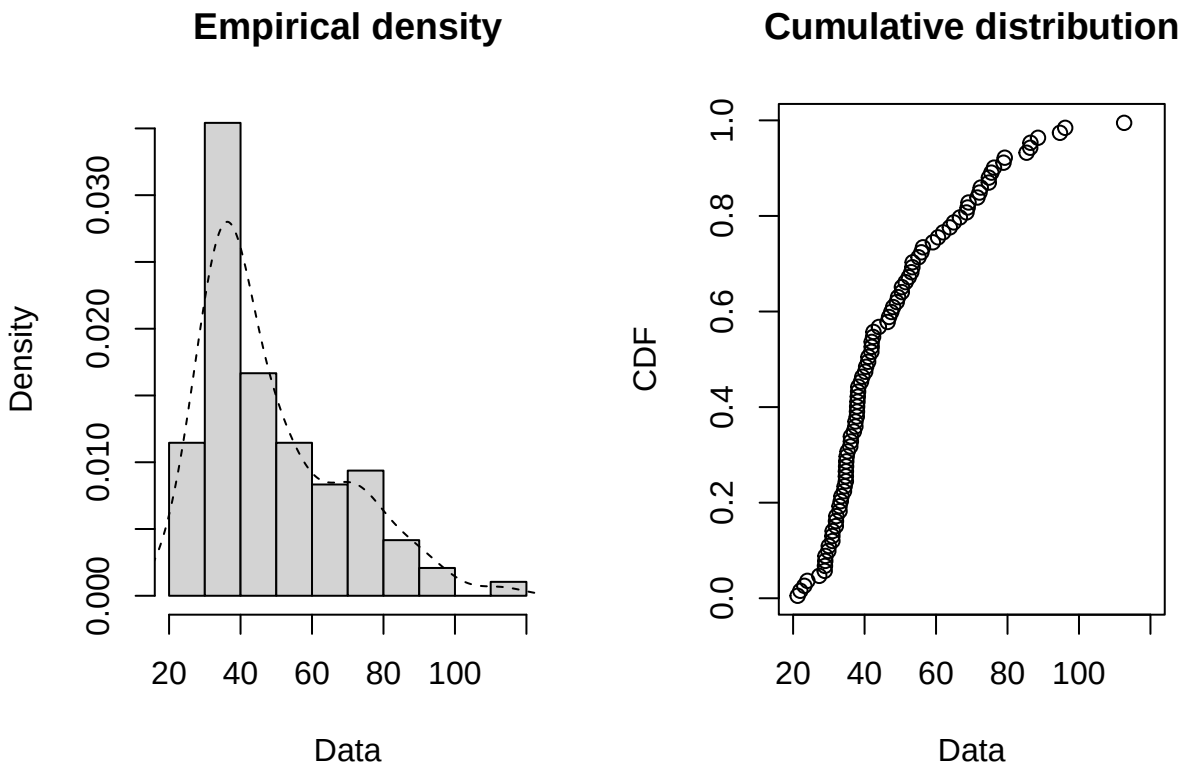


```
boxplot(data$price, main = "Price")
```

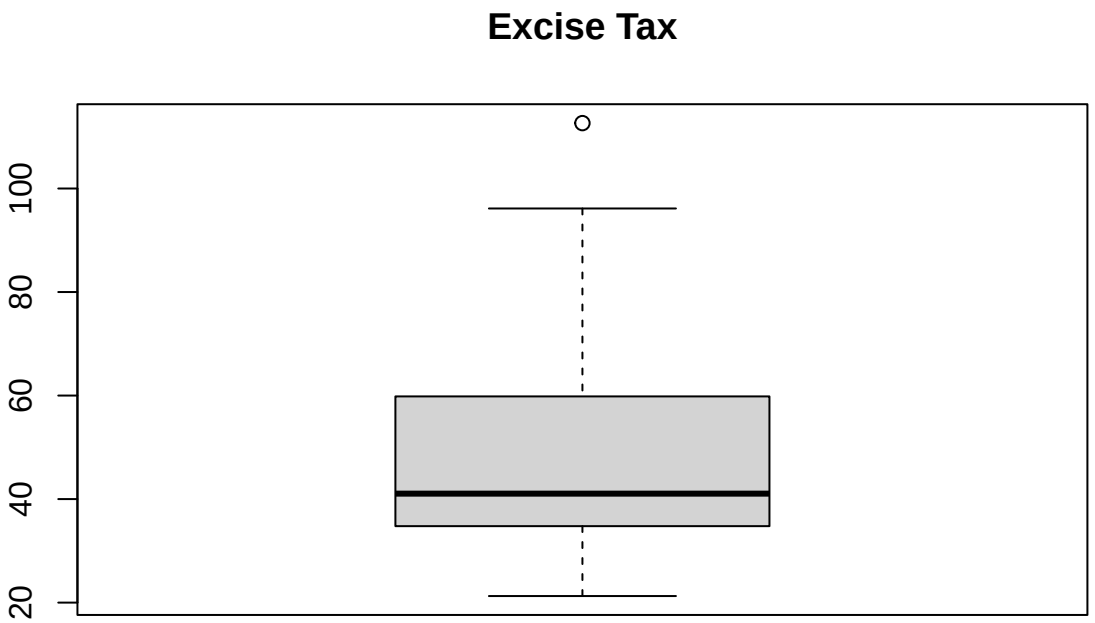


The histogram for cigarette price shows is skewed right, and is bimodal, with one mode around 100-115, and the other mean around 165-180. In addition, the box plot for price shows that there are no outliers.

```
# provide a histogram and box plot for tax
plotdist(data$taxs, histo = TRUE, demp = TRUE)
```

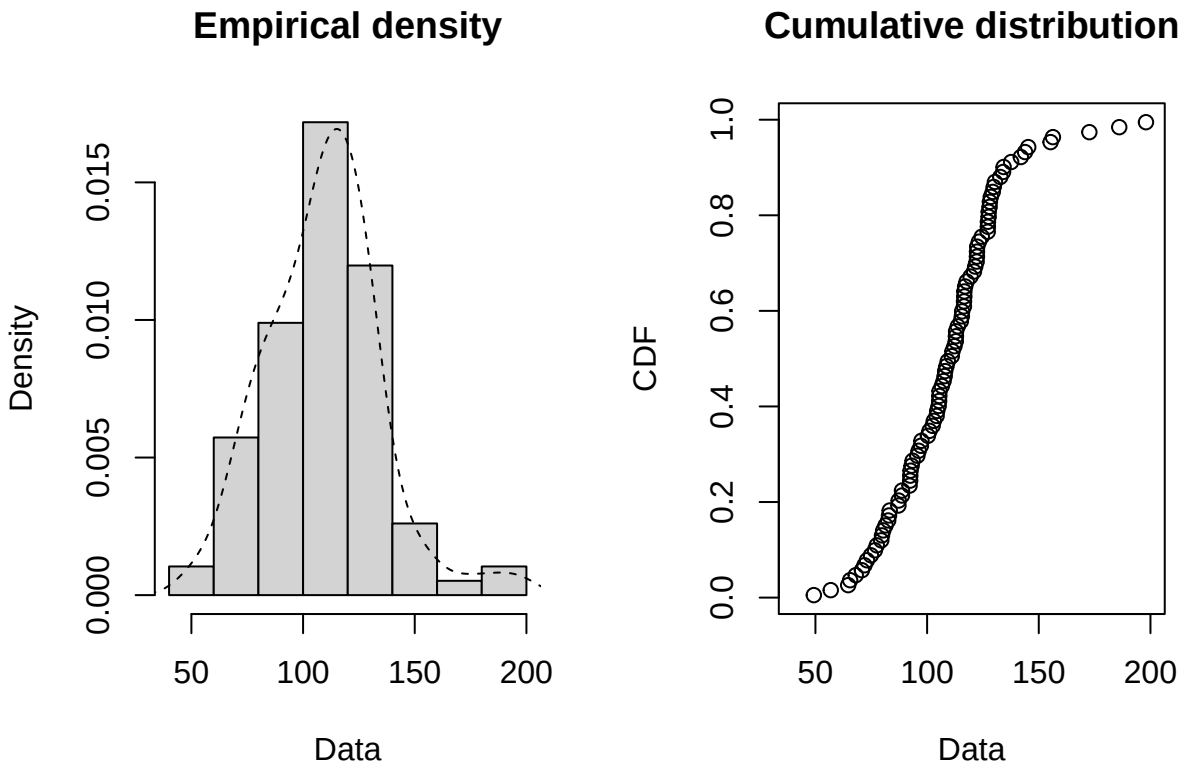


```
boxplot(data$taxs, main = "Excise Tax")
```

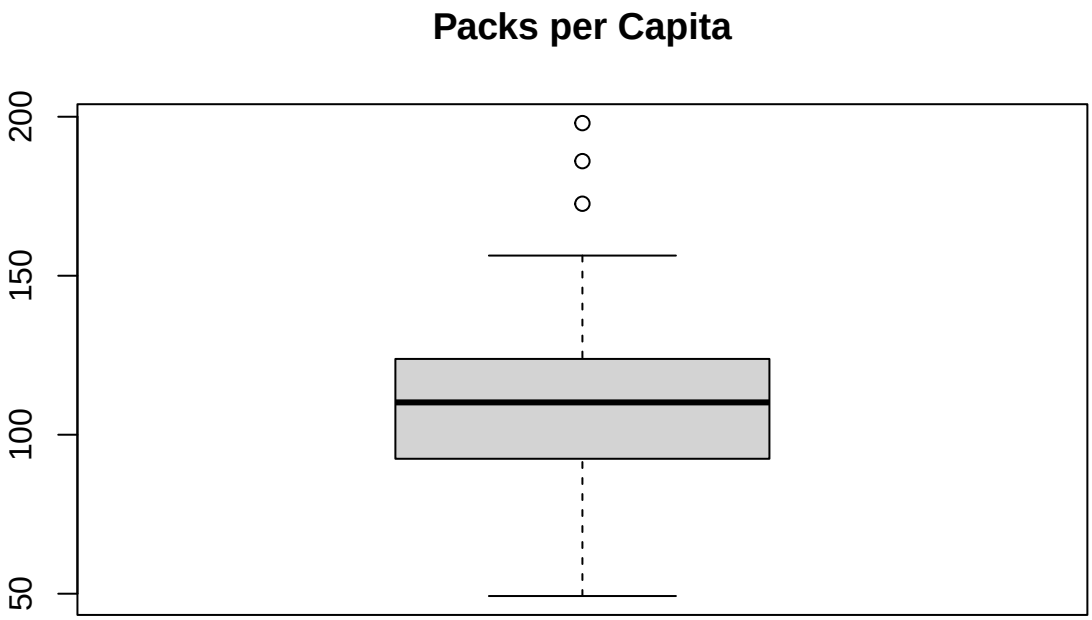


The histogram for excise taxes shows that it is skewed right, with the majority of its values around 30-40, and the box plot shows that there is one high outlier for excise taxes.

```
# provide a histograms and box plot for packs
plotdist(data$packs, histo = TRUE, demp = TRUE)
```

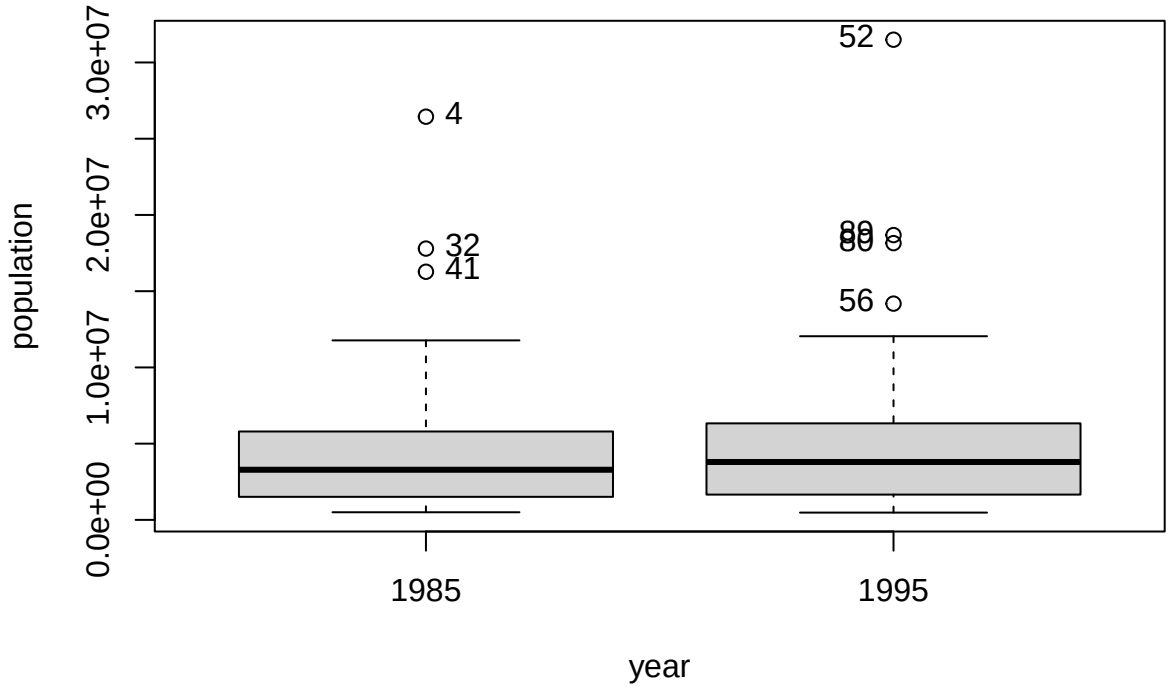


```
boxplot(data$packs, main = "Packs per Capita")
```



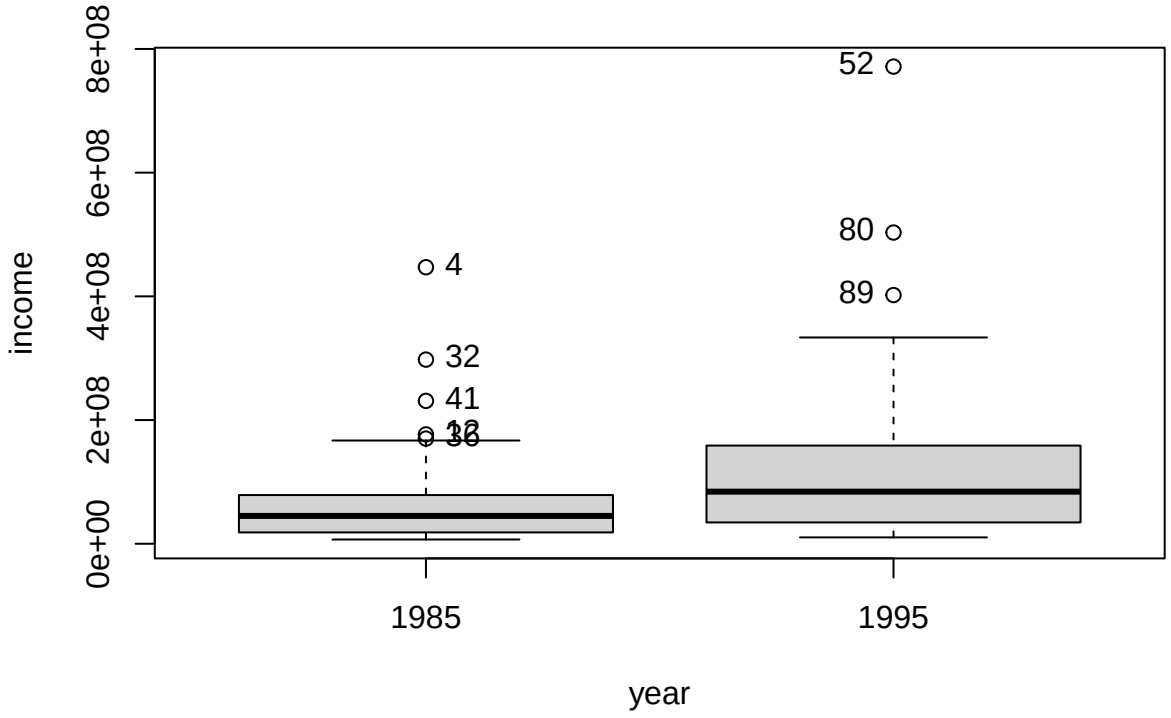
The histogram for packs shows a Normal distribution, with the mode around 100-115, and the box plot shows that there are 3 outliers on the higher side for packs per capita.

```
# provide boxplots for each of the variables by year
scatterplot(population ~ year | state, boxplots = TRUE, smooth = TRUE,
    data = data)
```



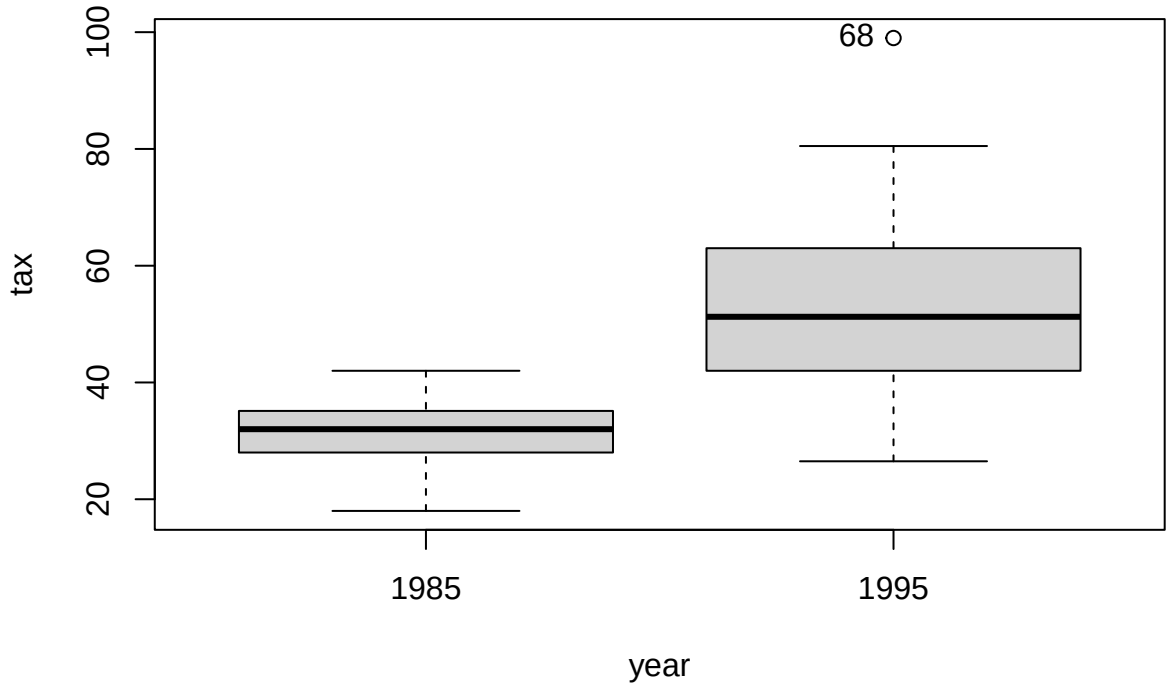

```
## [1] "4"  "32" "41" "52" "56" "80" "89"
```

```
scatterplot(income ~ year | state, boxplots = TRUE, smooth = TRUE,
    data = data)
```



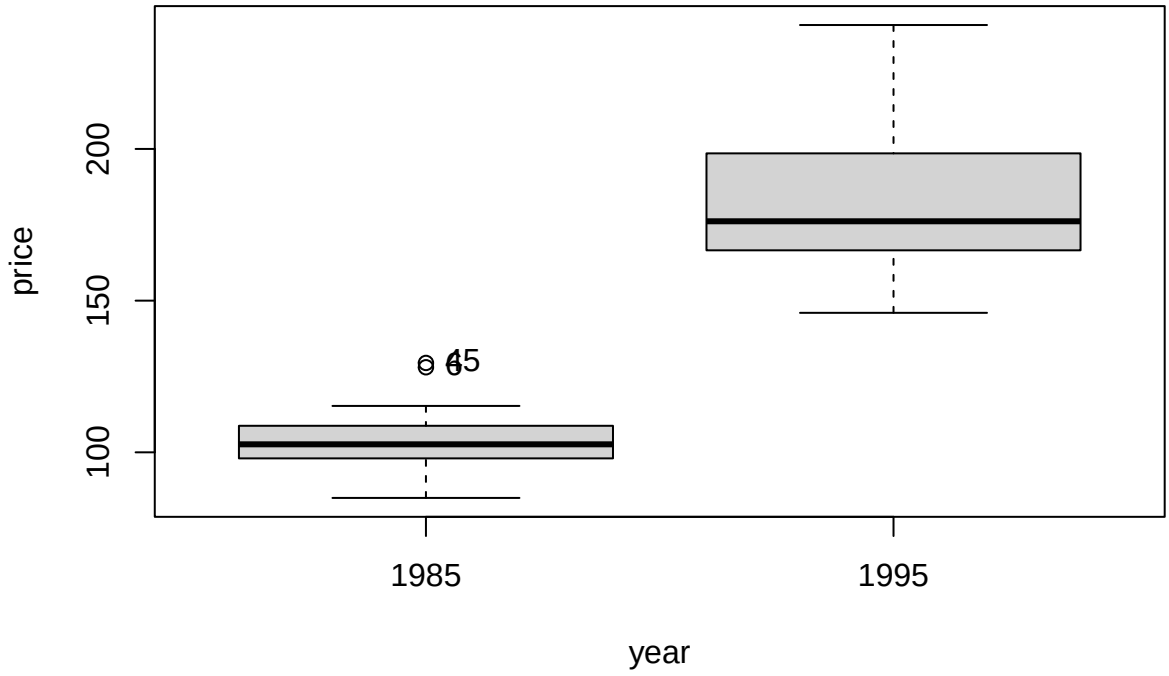

```
## [1] "4"  "12" "32" "36" "41" "52" "80" "89"
```

```
scatterplot(tax ~ year | state, boxplots = TRUE, smooth = TRUE,
    data = data)
```



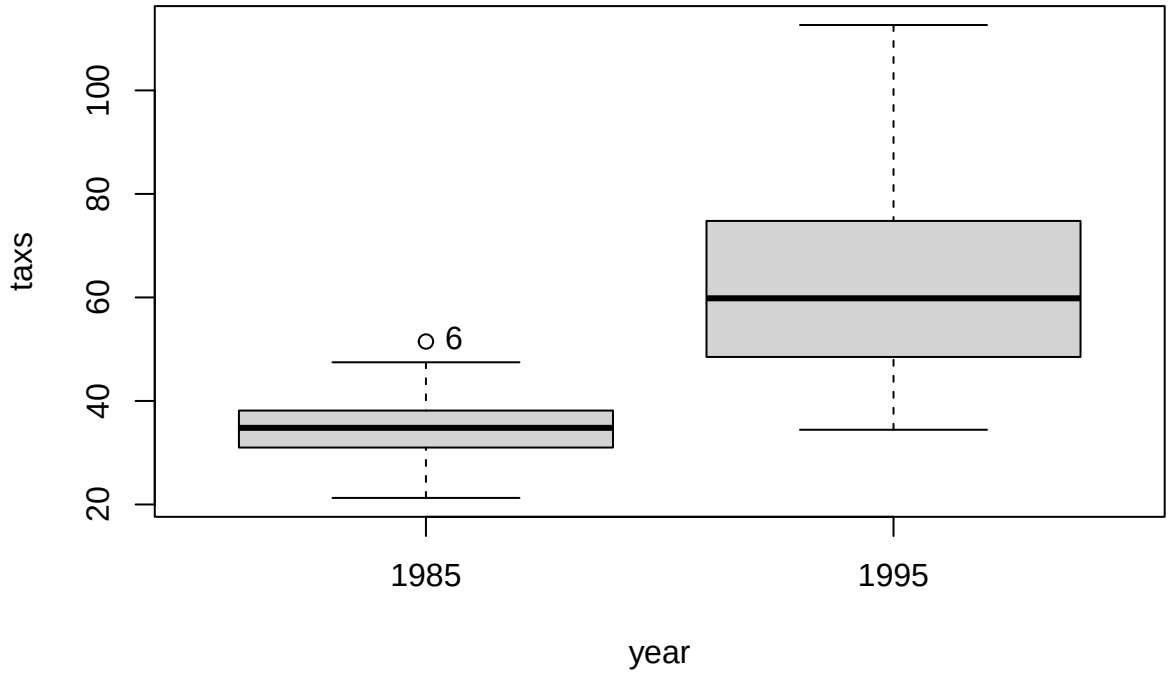
```
## [1] "68"
```

```
scatterplot(price ~ year | state, boxplots = TRUE, smooth = TRUE,
    data = data)
```



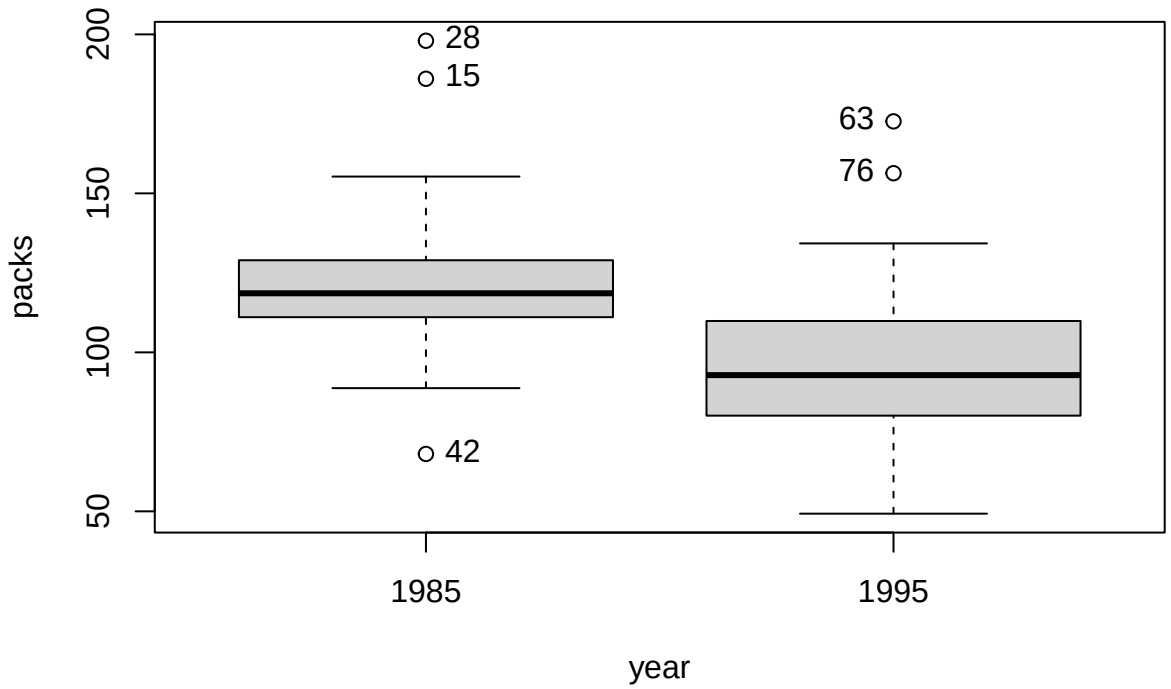

```
## [1] "6"  "45"
```

```
scatterplot(taxs ~ year | state, boxplots = TRUE, smooth = TRUE,
    data = data)
```



```
## [1] "6"
```

```
scatterplot(packs ~ year | state, boxplots = TRUE, smooth = TRUE,
    data = data)
```

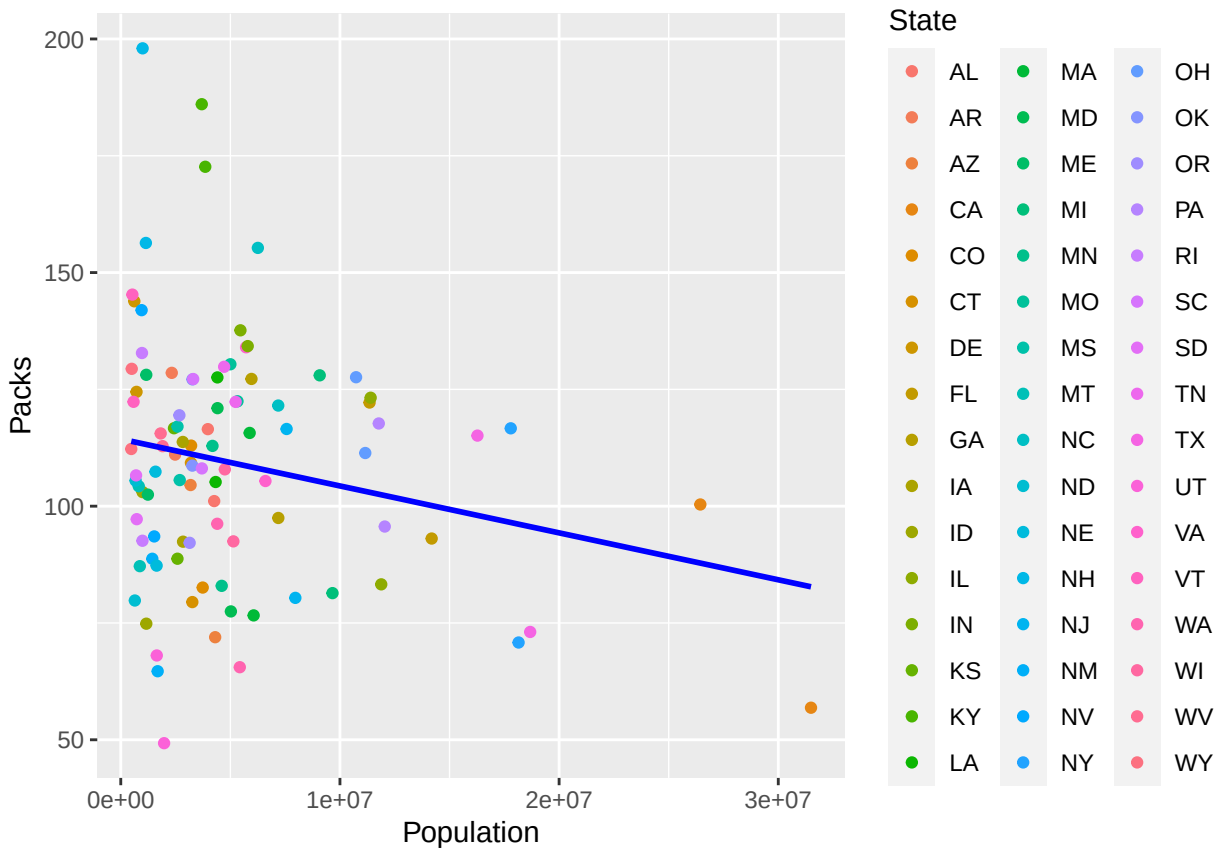



```
## [1] "42" "15" "28" "63" "76"
```

These boxplots are merely another way to show how data has changed over time from 1985 to 1995. All outliers appear to be natural and not due to error. For population, we can see there has been an increase in population over time. For income, we have seen there has been an increase in income over time. For excise taxes without sales tax, we have seen there has been an increase in excise taxes without sales tax over time. For price, we have seen there has been an incrcease in price over time. For excise taxes summed with sales tax, we have seen there has been an increase in excise taxes summed with sales tax over time. For number of packs per capita, we have seen there has been a decrease in the number of packs consumed per capita over time.

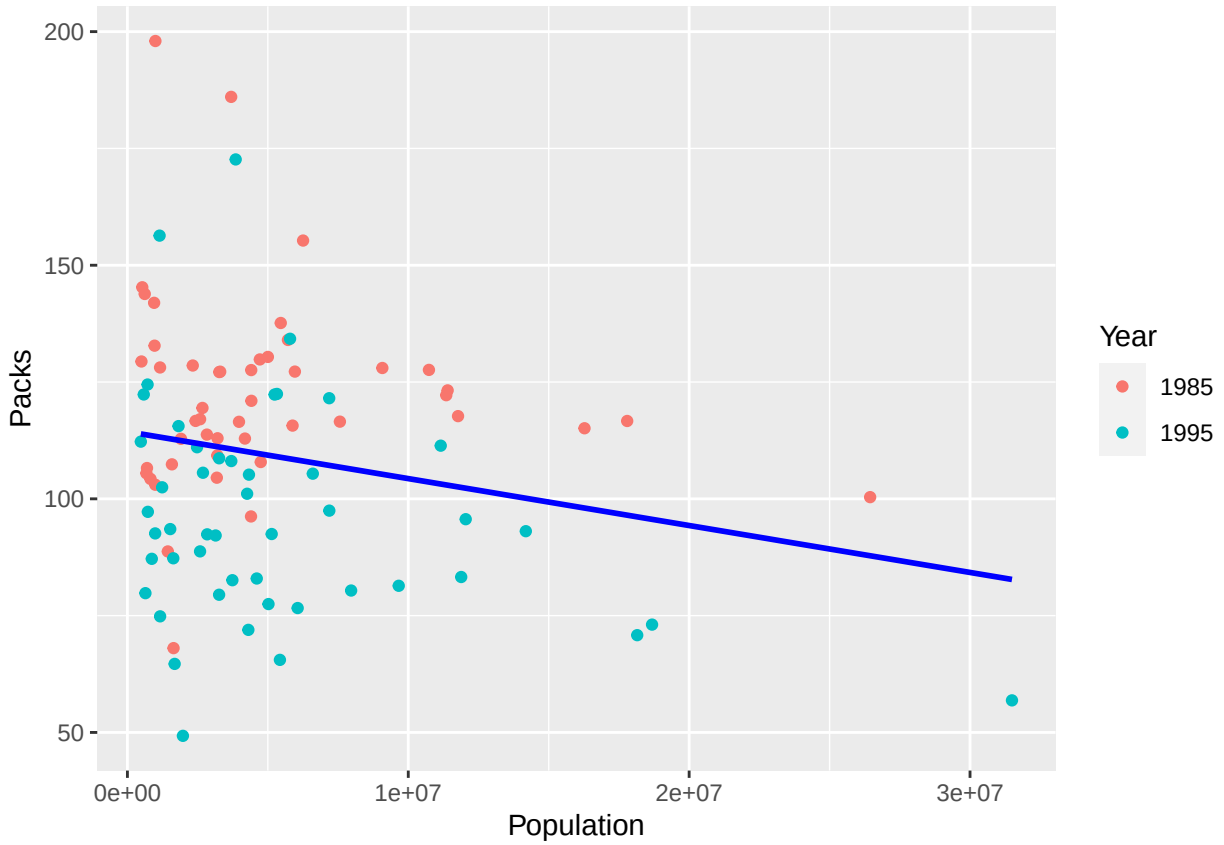
```
# provide a scatterplot of packs vs population by state
ggplot(data, aes(x = (population), y = (packs), colour = factor(state))) +
    geom_point() + xlab("Population") + ylab("Packs") + scale_colour_discrete(name = "State") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# provide a scatterplot of packs vs population by year
ggplot(data, aes(x = (population), y = (packs), colour = factor(year))) +
    geom_point() + xlab("Population") + ylab("Packs") + scale_colour_discrete(name = "Year") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

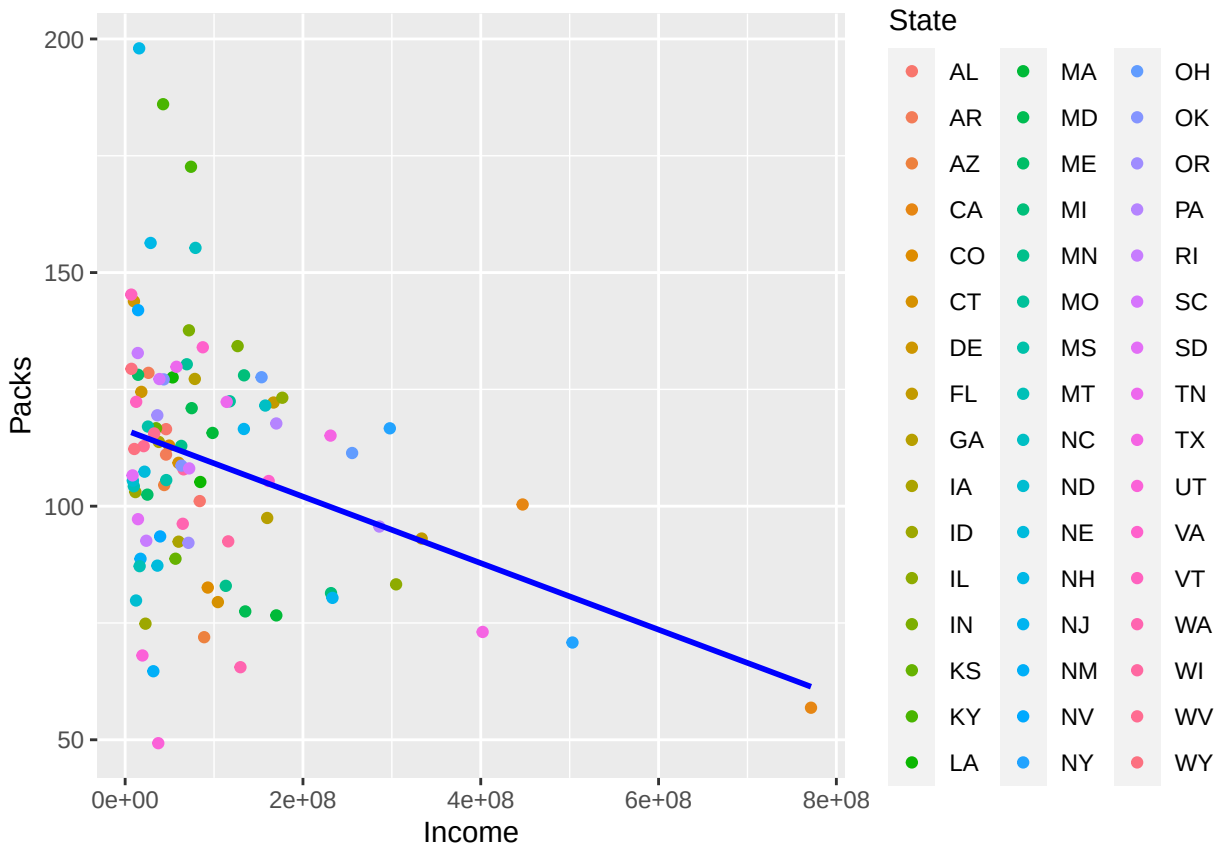
```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplots of packs against population by state and year have a weak, negative relationship. It doesn't appear as though population or packs per capita is dependent on the state, however it does appear that in 1995, the number of packs per capita consumed decreased.

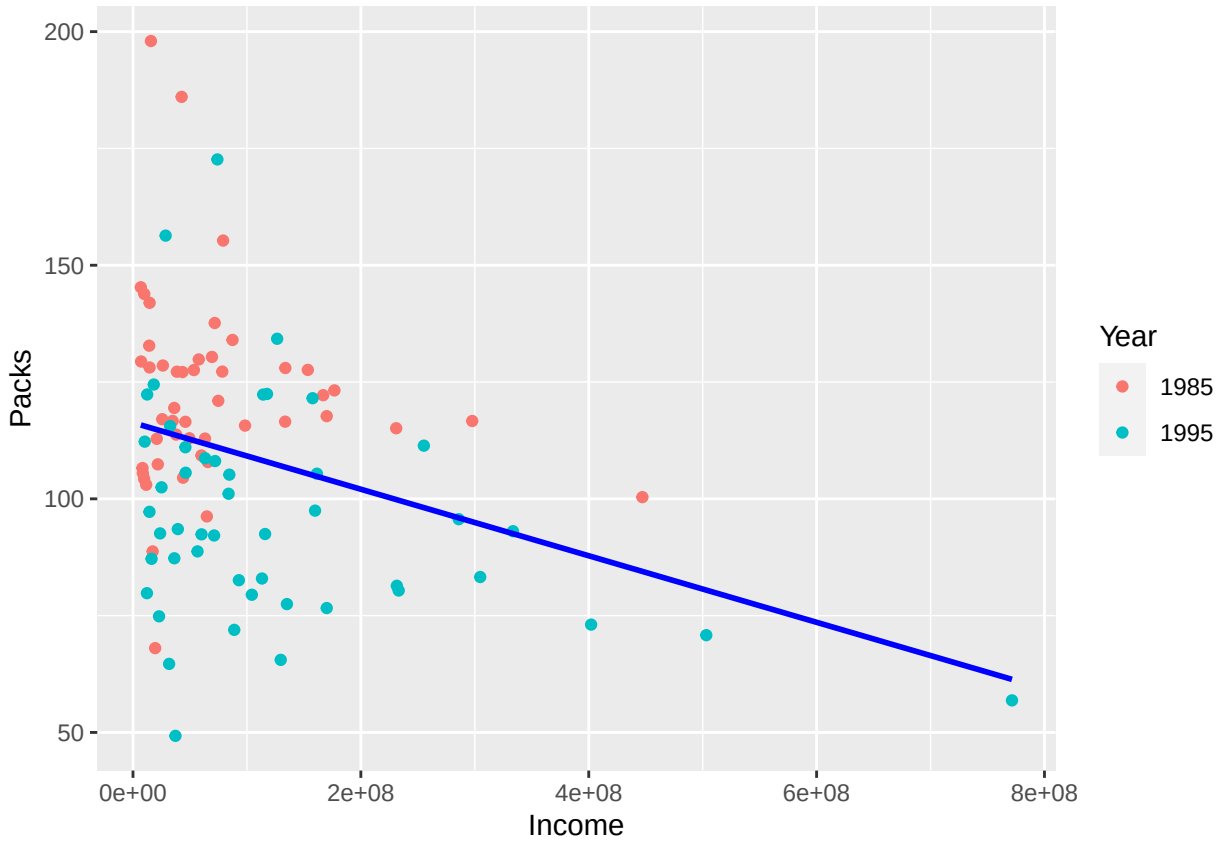
```
# provide a scatterplot of packs vs income by state
ggplot(data, aes(x = (income), y = (packs), colour = factor(state))) +
    geom_point() + xlab("Income") + ylab("Packs") + scale_colour_discrete(name = "State") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# provide a scatterplot of packs vs income by year
ggplot(data, aes(x = (income), y = (packs), colour = factor(year))) +
    geom_point() + xlab("Income") + ylab("Packs") + scale_colour_discrete(name = "Year") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

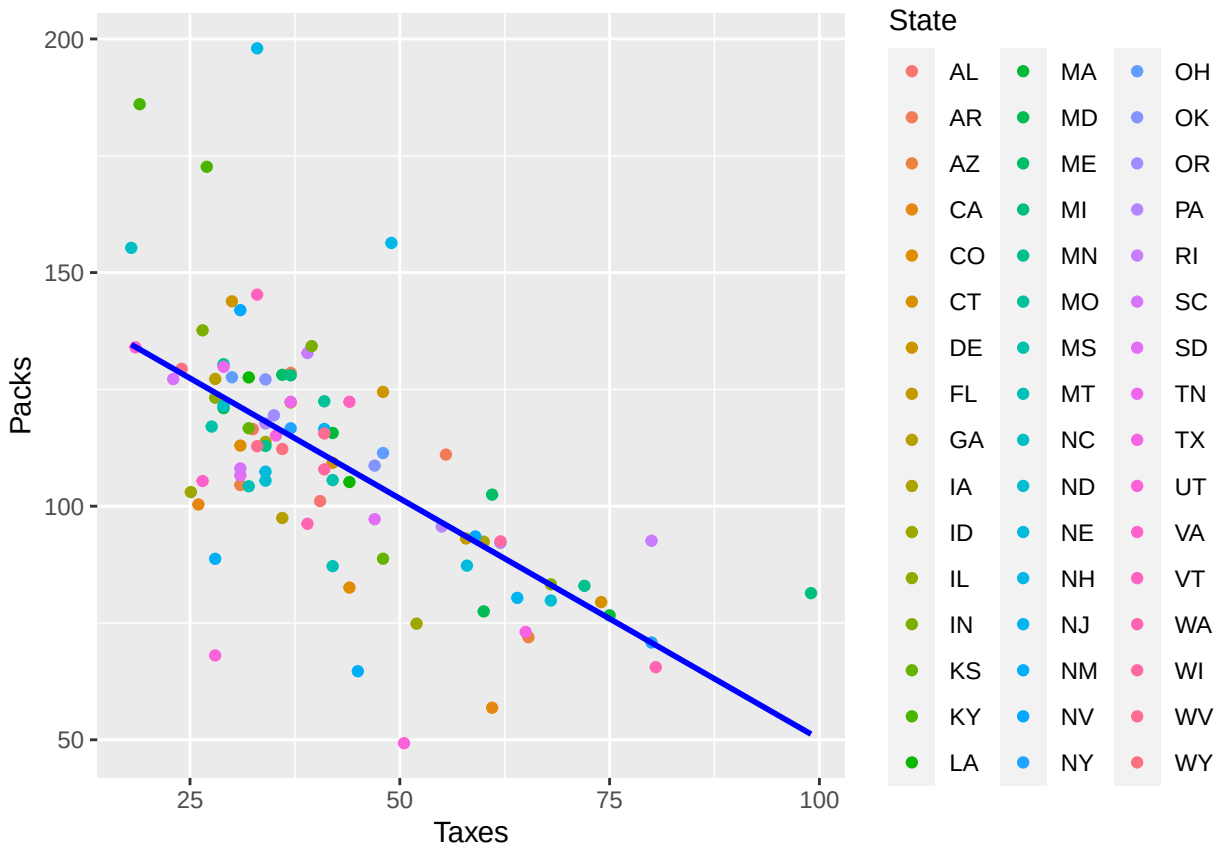
```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplot of packs per capita against income by state and year have a weak, relationship between the two, similarly to the scatterplot of packs against population. Also similarly to the previous scatterplot, it appears as the packs per capita are lower in 1995 than in 1985, regardless of income or state.

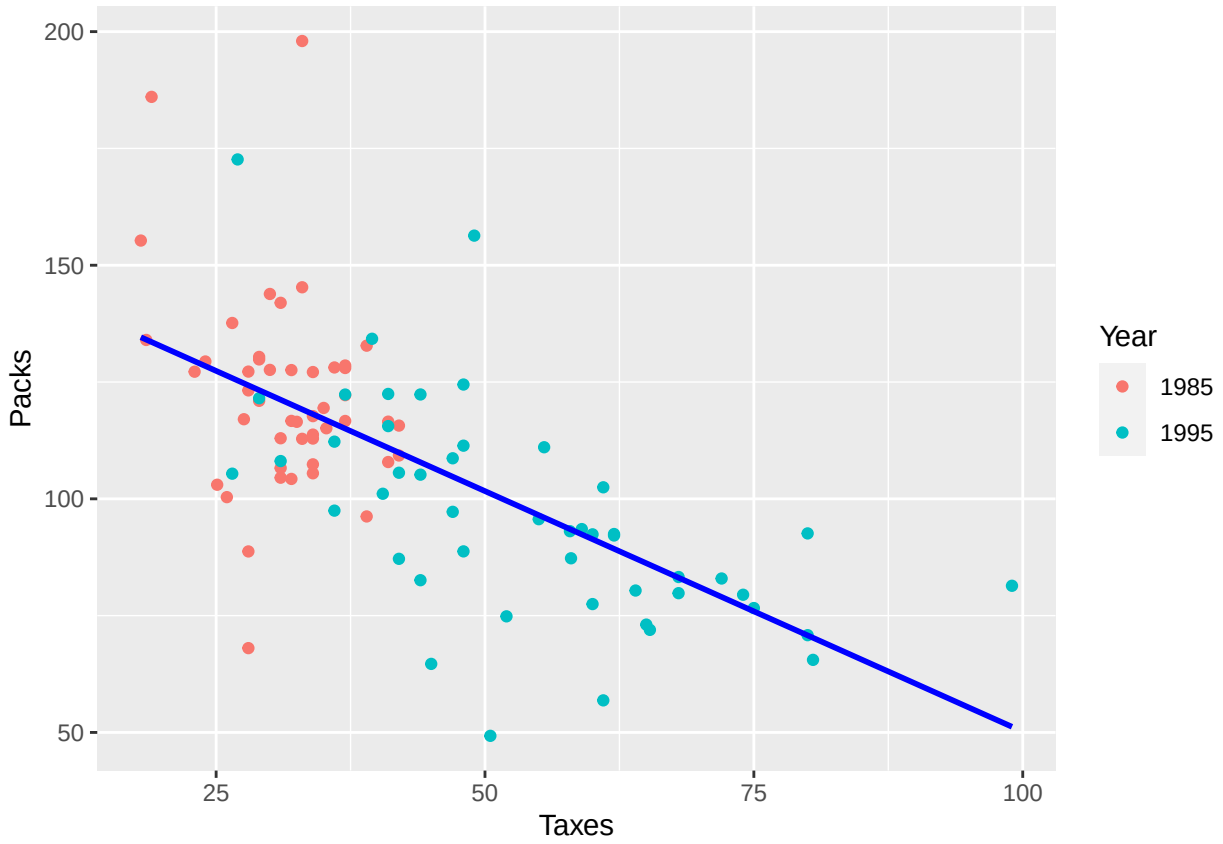
```
# provide a scatterplot of packs vs taxes by state
ggplot(data, aes(x = (tax), y = (packs), colour = factor(state))) +
    geom_point() + xlab("Taxes") + ylab("Packs") + scale_colour_discrete(name = "State") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# provide a scatterplot of packs vs taxes by year
ggplot(data, aes(x = (tax), y = (packs), colour = factor(year))) +
    geom_point() + xlab("Taxes") + ylab("Packs") + scale_colour_discrete(name = "Year") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

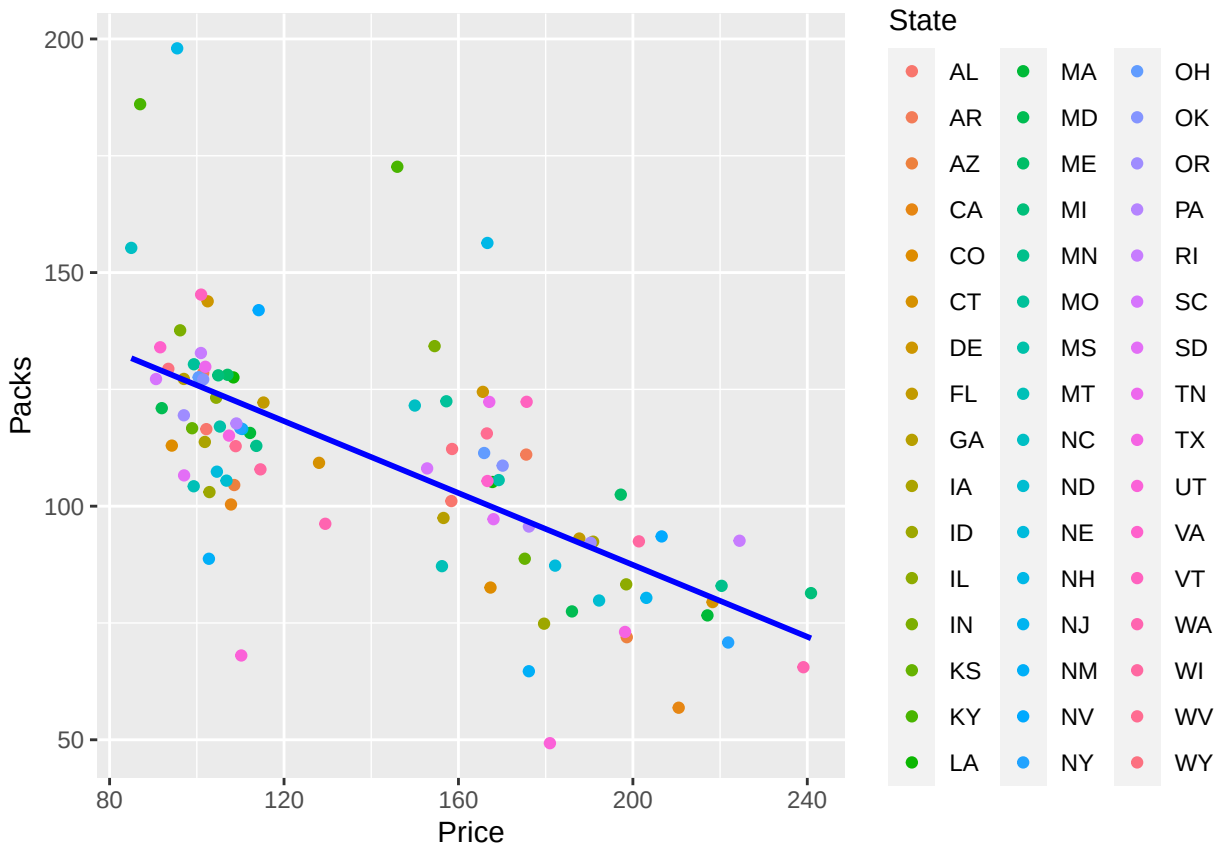
```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplots of packs against taxes by state and year show a strong, negative relationship between them, so when taxes are higher, cigarette consumption is lower, and vice-versa. In addition, it doesn't appear as though there is a connection between state and the relationship between taxes and packs per capita, but in 1985 it appears that consumption was higher and taxes were lower, and in 1995 consumption was lower and taxes were higher.

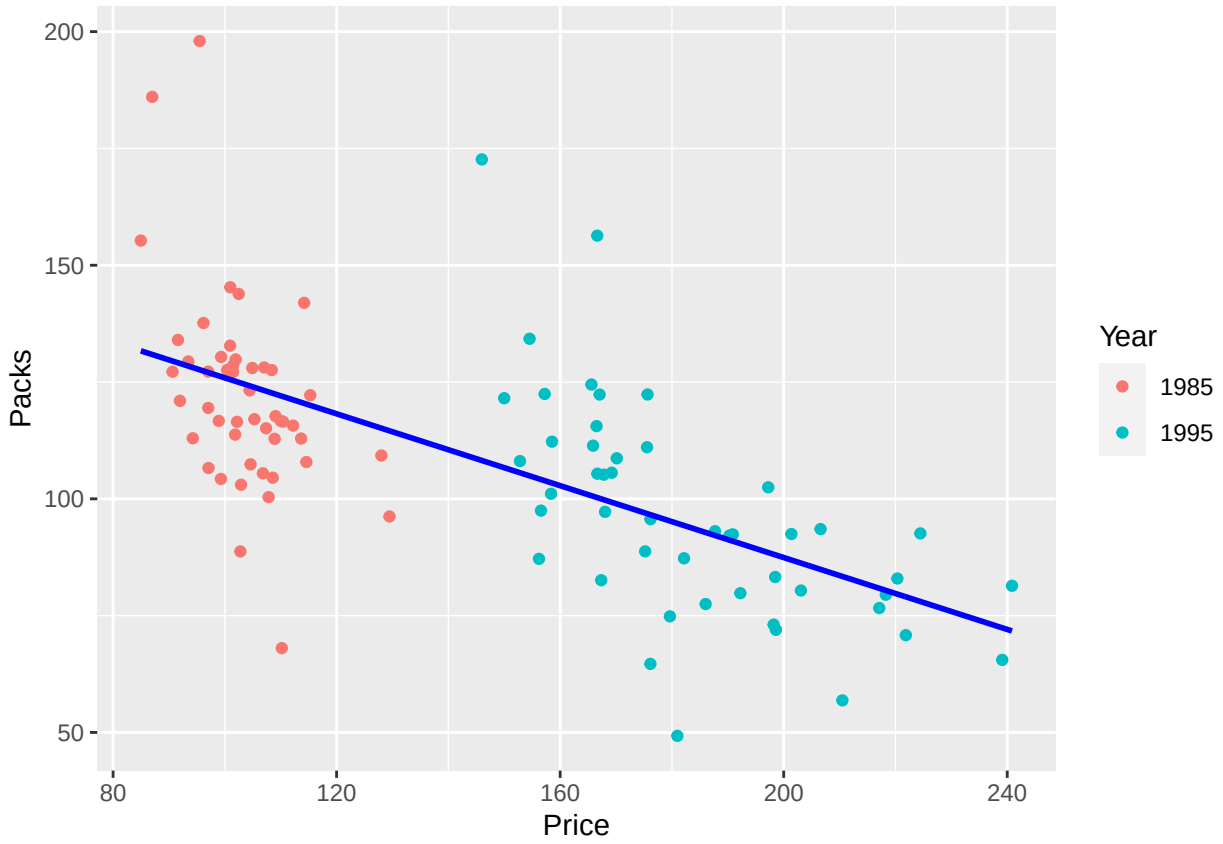
```
# provide a scatterplot of packs vs price by state
ggplot(data, aes(x = (price), y = (packs), colour = factor(state))) +
    geom_point() + xlab("Price") + ylab("Packs") + scale_colour_discrete(name = "State") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# provide a scatterplot of packs vs price by year
ggplot(data, aes(x = (price), y = (packs), colour = factor(year))) +
    geom_point() + xlab("Price") + ylab("Packs") + scale_colour_discrete(name = "Year") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

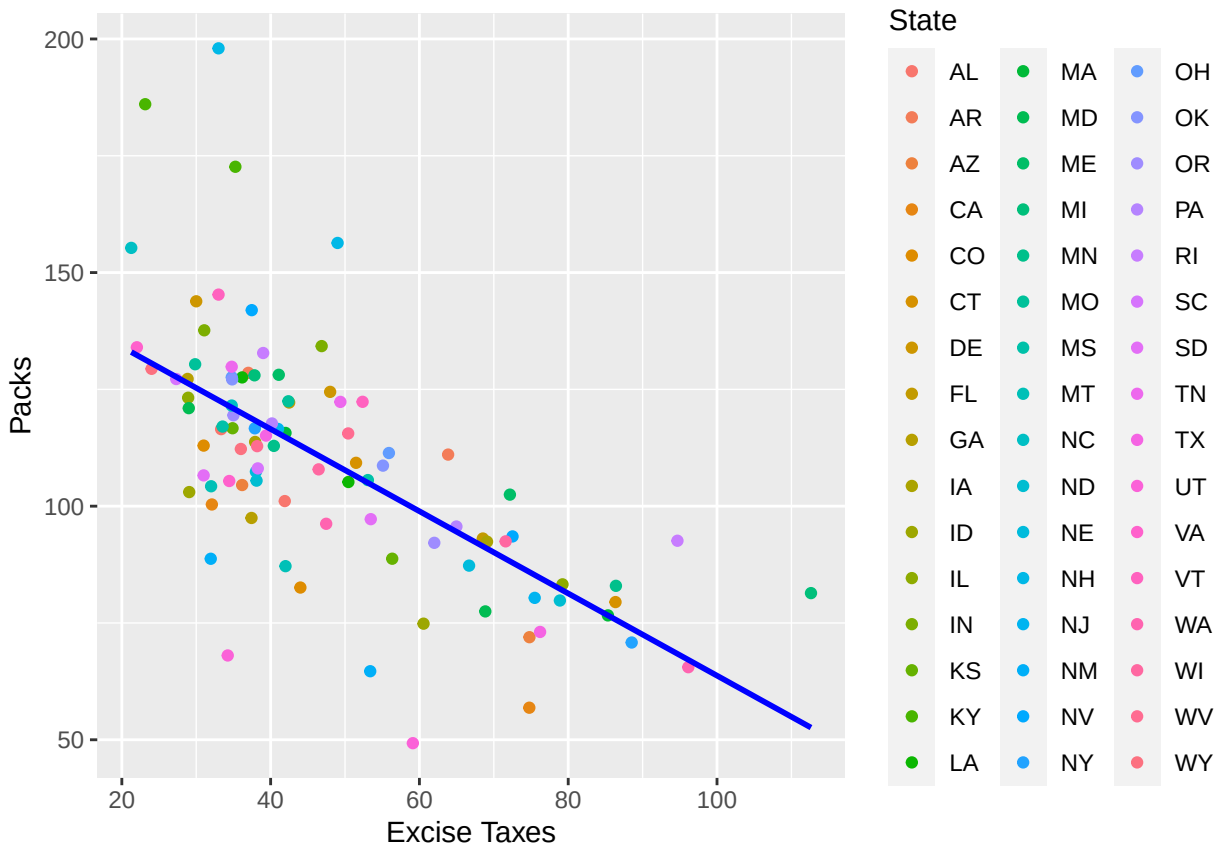
```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplots of consumption of cigarette packs per capita against the price of a pack of cigarettes show a strong, negative correlation, so when price is lower consumption is higher, and vice-versa. In addition, there doesn't appear to be a relationship between state and packs vs price, however, similarly to with packs vs taxes, in 1985 consumption was higher and prices were lower, and vice versa in 1995.

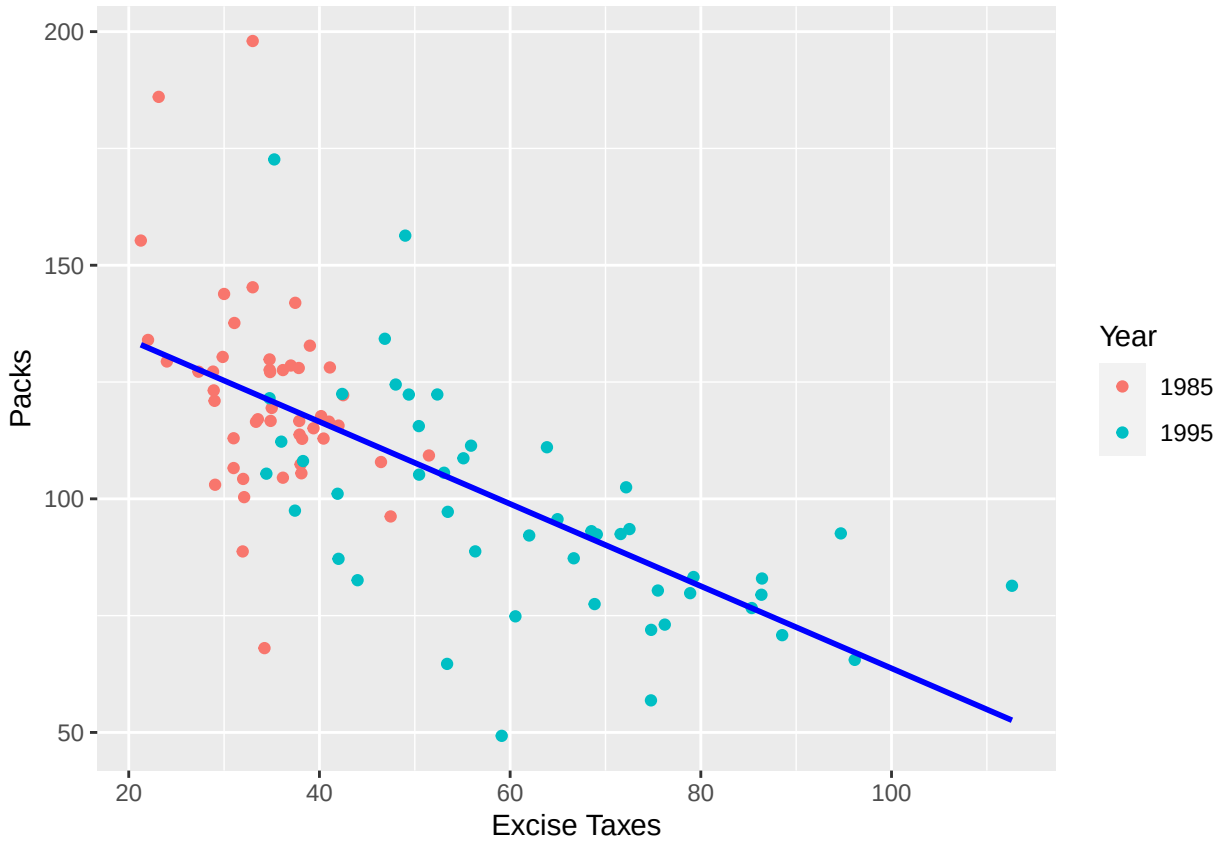
```
# provide a scatterplot of packs vs excise taxes by state
ggplot(data, aes(x = (taxs), y = (packs), colour = factor(state))) +
    geom_point() + xlab("Excise Taxes") + ylab("Packs") + scale_colour_discrete(name = "State") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



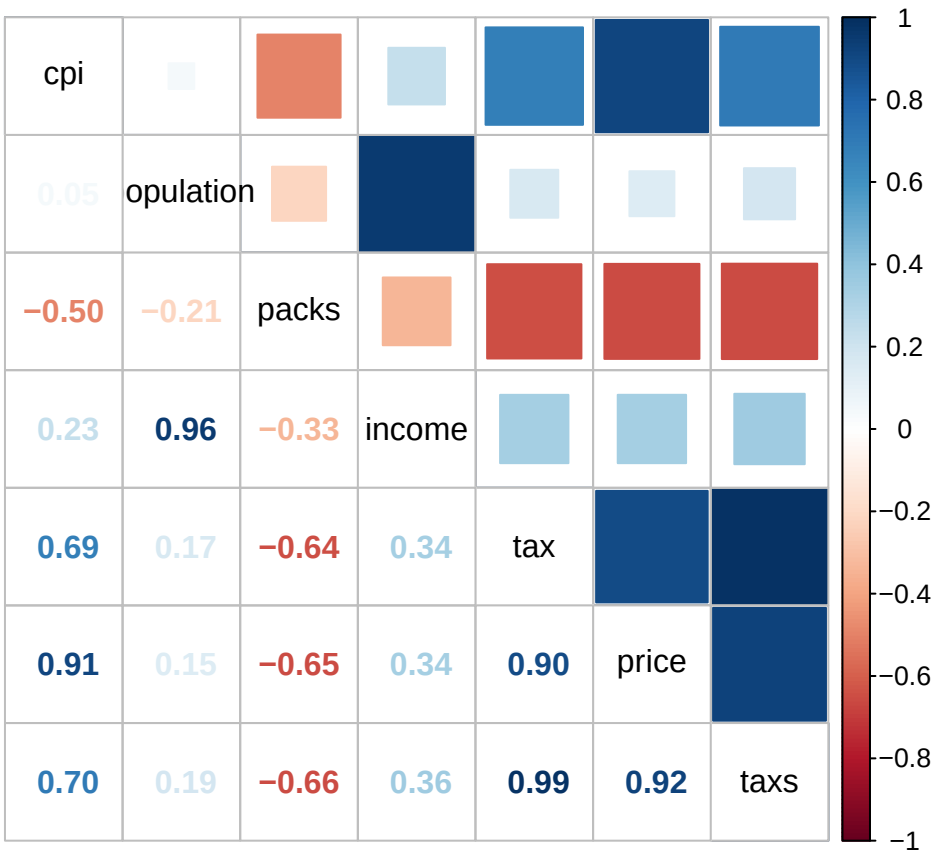

```
# provide a scatterplot of packs vs excise taxes by year
ggplot(data, aes(x = (taxs), y = (packs), colour = factor(year))) +
    geom_point() + xlab("Excise Taxes") + ylab("Packs") + scale_colour_discrete(name = "Year") +
    geom_smooth(method = lm, color = "blue", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Similarly to with packs against taxes and packs against price, the scatter plots of packs against excise taxes shows a strong, negative correlation, which seems unaffected by state, but shows higher consumption and lower excise tax in 1985, and the opposite in 1995.

```
# provide a correlation plot of the variables
corrdata <- data.frame(cpi, population, packs, income, tax, price,
    taxs)
corrplot.mixed(cor(corrdata), upper = "square", lower = "number",
    addgrid.col = "black", tl.col = "black")
```



The correlation plot above shows the correlation between each of the variables in our model. We can see that population and income, tax and taxs, tax and price, and cpi and price have strong positive correlations with values all greater than or equal to 0.9, while packs has a negative relationship with all of the explanatory variables.

Because the explanatory "tax" and "taxs", as well as "price" and "taxs" have such a high, strong, positive correlation, we can remove the variable "taxs" from our data set, as it may effect the results of our models. In addition, because "cpi" and "price" have such a strong positive correlation, we can remove the variable "cpi" from the data as well.

```
data <- data.frame(state, year, population, packs, income, tax,
    price)
```

(3) Fit a pooled model, a fixed effects model, and a random effects model to conclude which one is preferred for the data. Include other evidence (plots, statistical diagnostics, etc.) to support the conclusion. For each of the models, because the distributions of each of the explanatory variables is not Normal, use the log.

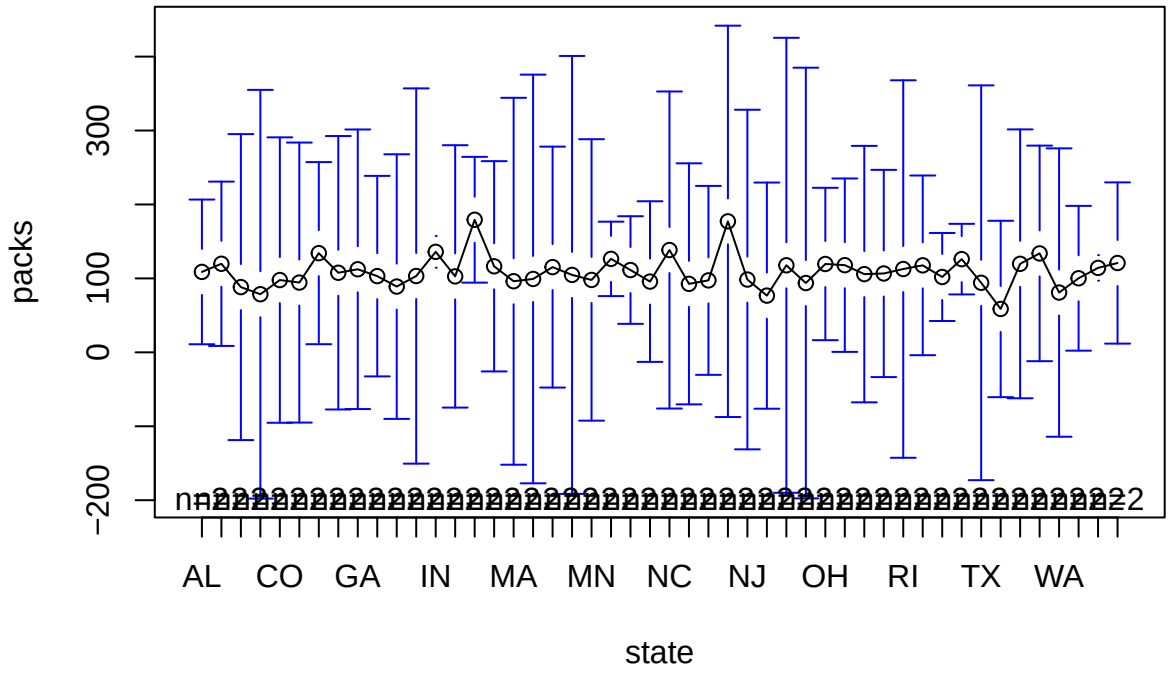
```
# provide visual evidence to see if heterogeneity exists
# across airline firms or years
plotmeans(packs ~ state, data = data)
```

```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```

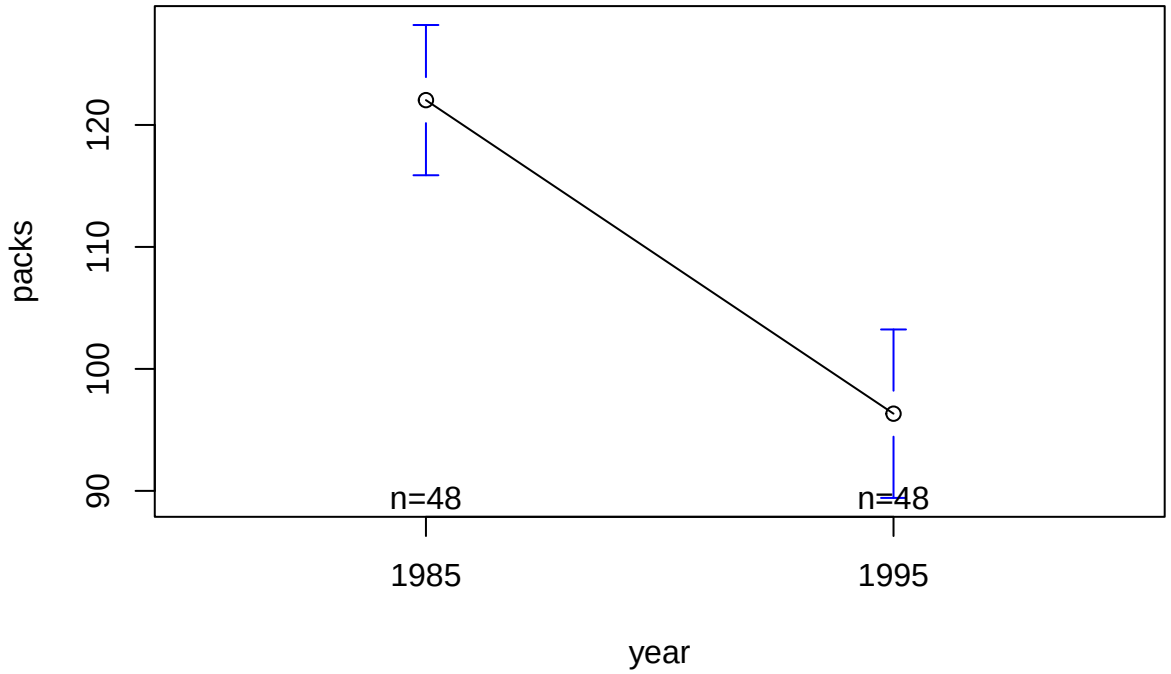
```
## Warning in arrows(x, li, x, pmax(y - gap, li), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped

## Warning in arrows(x, ui, x, pmin(y + gap, ui), col = barcol, lwd = lwd, : zero-
## length arrow is of indeterminate angle and so skipped
```




```
plotmeans(packs ~ year, data = data)
```



From the plots of the means over time and individuals show that individual and time heterogeneity both exist in our model, as the means for both vary.

```
# perform a Breusch-Pagan Test to test for
# heteroskedasticity
bptest(packs ~ log(population) + log(income) + log(tax) + log(price),
    data = data)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  packs ~ log(population) + log(income) + log(tax) + log(price)
## BP = 8.7738, df = 4, p-value = 0.06701
```

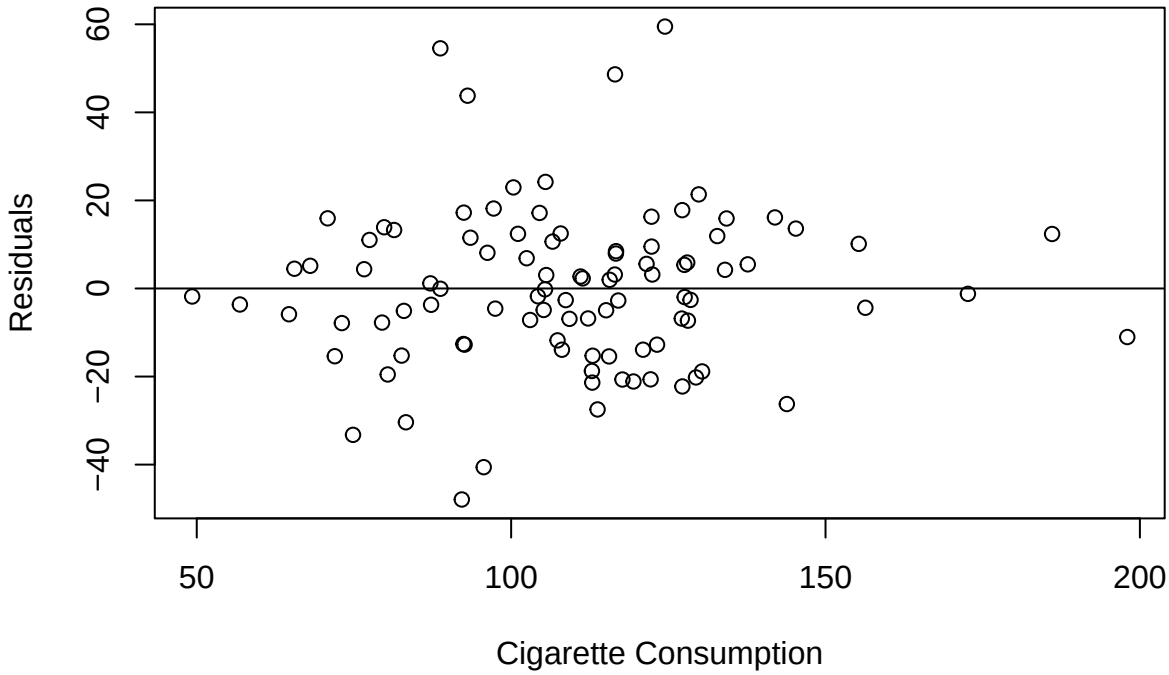
The p-value of our BP test is 0.0671, which is greater than 0.05, therefore we can fail to reject the null that our errors are uncorrelated, and conclude that heteroskedasticity does not exist in our data set.

When fitting a pooled model, we use the following hypotheses; null hypothesis: pooled model is appropriate, alternative hypothesis: pooled model is not appropriate, and we reject the null hypothesis when the p-value is less than 0.05.

```
# fit a pooled model for the data
pooled <- plm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price), data = data, model = "pooling")
summary(pooled)
```

```
## Pooling Model
##
## Call:
## plm(formula = packs ~ log(population) + log(income) + log(tax) +
##     log(price), data = data, model = "pooling")
##
## Balanced Panel: n = 48, T = 2, N = 96
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -47.92253 -11.99077  -0.70922  10.73740  59.48371
##
## Coefficients:
##                 Estimate Std. Error t-value  Pr(>|t|)
## (Intercept)      454.520     49.688  9.1476 1.551e-14 ***
## log(population)  -53.854     15.882 -3.3908 0.0010324 **
## log(income)       50.095     15.273  3.2800 0.0014719 **
## log(tax)         -23.522     11.094 -2.1201 0.0367148 *
## log(price)       -70.222     18.763 -3.7425 0.0003187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    63586
## Residual Sum of Squares: 30670
## R-Squared:      0.51766
## Adj. R-Squared: 0.49645
## F-statistic: 24.4155 on 4 and 91 DF, p-value: 9.6126e-14
```

```
# provide a residual plot for the pooled model against
# cigarette consumption
pooled.res <- resid(pooled)
plot(packs, pooled.res, ylab = "Residuals", xlab = "Cigarette Consumption")
abline(0, 0)
```



We can see from our regression above that the p-value is 9.6126e-14, which is less than 0.05, and can conclude that a pooled model is not appropriate for our data because it doesn't consider heterogeneity across states or years. We can also see from the regression above that the p-values for all of the response variables are significant because they are all less than 0.05. In addition, the plot of the residuals of the pooled model are not noisy and show that there may be a slight relationship between the residual errors and the response variable.

When fitting a least squares dummy variable fixed model, we use the null hypothesis that all dummy parameters except one is 0, and we reject the null hypothesis when the p-value is less than 0.05.

```
# fit a least square dummy variable fixed effects model for
# the data
lsdv <- lm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price) + factor(state) - 1, data = data)
summary(lsdv)
```

```
##
## Call:
## lm(formula = packs ~ log(population) + log(income) + log(tax) +
##     log(price) + factor(state) - 1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.322 -2.329  0.000  2.329  7.322
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## log(population)  -86.800     19.354  -4.485 5.17e-05 ***
## log(income)       46.116     13.683   3.370 0.001571 **
## log(tax)         -10.880      9.299  -1.170 0.248316
## log(price)       -68.603     17.197  -3.989 0.000247 ***
## factor(state)AL  974.735    172.489   5.651 1.10e-06 ***
## factor(state)AR  971.566    166.121   5.849 5.63e-07 ***
## factor(state)AZ  956.948    170.829   5.602 1.29e-06 ***
## factor(state)CA 1022.655    193.946   5.273 3.90e-06 ***
## factor(state)CO  944.045    169.405   5.573 1.43e-06 ***
## factor(state)CT  951.627    167.760   5.673 1.02e-06 ***
## factor(state)DE  914.740    150.569   6.075 2.62e-07 ***
## factor(state)FL 1022.409    184.416   5.544 1.57e-06 ***
## factor(state)GA  987.855    177.530   5.564 1.47e-06 ***
## factor(state)IA  957.501    167.575   5.714 8.87e-07 ***
## factor(state)ID  904.760    157.538   5.743 8.04e-07 ***
## factor(state)IL 1009.161    183.241   5.507 1.78e-06 ***
## factor(state)IN 1005.119    175.735   5.720 8.70e-07 ***
## factor(state)KS  944.330    166.129   5.684 9.80e-07 ***
## factor(state)KY 1028.938    172.403   5.968 3.76e-07 ***
## factor(state)LA  988.239    173.240   5.704 9.16e-07 ***
## factor(state)MA  979.384    174.947   5.598 1.31e-06 ***
## factor(state)MD  958.094    172.414   5.557 1.50e-06 ***
## factor(state)ME  940.536    158.092   5.949 4.01e-07 ***
## factor(state)MI 1014.850    180.491   5.623 1.21e-06 ***
## factor(state)MN  973.629    172.282   5.651 1.10e-06 ***
## factor(state)MO  992.984    174.551   5.689 9.65e-07 ***
## factor(state)MS  968.487    168.440   5.750 7.86e-07 ***
## factor(state)MT  896.376    154.410   5.805 6.52e-07 ***
## factor(state)NC 1006.462    178.478   5.639 1.14e-06 ***
## factor(state)ND  893.684    151.234   5.909 4.59e-07 ***
## factor(state)NE  926.187    160.991   5.753 7.77e-07 ***
## factor(state)NH  976.104    155.722   6.268 1.36e-07 ***
## factor(state)NJ  986.287    177.851   5.546 1.56e-06 ***
## factor(state)NM  906.884    161.752   5.607 1.27e-06 ***
## factor(state)NV  935.709    157.715   5.933 4.23e-07 ***
## factor(state)NY 1021.781    187.636   5.446 2.19e-06 ***
## factor(state)OH 1018.386    182.864   5.569 1.44e-06 ***
## factor(state)OK  974.977    169.486   5.753 7.79e-07 ***
## factor(state)OR  957.863    167.477   5.719 8.71e-07 ***
## factor(state)PA 1014.237    183.660   5.522 1.69e-06 ***
## factor(state)RI  926.282    154.742   5.986 3.54e-07 ***
## factor(state)SC  968.379    171.192   5.657 1.08e-06 ***
## factor(state)SD  896.452    152.349   5.884 4.99e-07 ***
## factor(state)TN  996.719    174.718   5.705 9.15e-07 ***
## factor(state)TX 1024.440    188.155   5.445 2.19e-06 ***
## factor(state)UT  899.039    163.492   5.499 1.83e-06 ***
## factor(state)VA  983.176    177.362   5.543 1.58e-06 ***
## factor(state)VT  917.446    149.378   6.142 2.09e-07 ***
## factor(state)WA  970.808    173.702   5.589 1.35e-06 ***
## factor(state)WI  982.020    173.747   5.652 1.09e-06 ***
## factor(state)WV  955.476    164.004   5.826 6.08e-07 ***
## factor(state)WY  888.000    148.087   5.996 3.42e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.082 on 44 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.9979
## F-statistic: 898.5 on 52 and 44 DF,  p-value: < 2.2e-16
```

In this model we used each state as a dummy variable. We can see from our LSDV model regression that our p-value is <2.2e-16, so we can reject our null hypothesis and conclude that this model is better than our pooled model. We can also see that the individual effects for all of the variables are insignificant, except for the dummy variables for all of the states and price.

When fitting a within fixed effects model, we use the same null hypothesis that we used for the LSDV fixed model, and we reject the null hypothesis when our p-value is less than 0.05.

```
# fit an individual effect, time effect, and two way effect
# within fixed effects model for the data case 1:
# individual effect model
fm_ind <- plm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price), data = data, model = "within", effect = "individual")

# case 2: time effect model
fm_time <- plm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price), data = data, model = "within", effect = "time")

# case 3: two way effect model
fm_both <- plm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price), data = data, model = "within", effect = "twoways")

stargazer(fm_ind, fm_time, fm_both, type = "text")
```

```
##
## ===============================================================================
##                                      Dependent variable:
##                  ---------------------------------------------------------------
##                                             packs
##                           (1)                 (2)                  (3)
## -------------------------------------------------------------------------------
## log(population)       -86.800***          -41.914**            -58.491**
##                        (19.354)            (16.439)             (27.916)
##
## log(income)            46.116***           39.503**             17.654
##                        (13.683)            (15.683)             (24.502)
##
## log(tax)               -10.880             -1.942               -9.475
##                        (9.299)             (14.536)             (9.256)
##
## log(price)            -68.603***          -135.260***          -80.163***
##                        (17.197)            (34.430)             (18.930)
##
## -------------------------------------------------------------------------------
## Observations              96                  96                   96
## R2                       0.941               0.391                0.681
## Adjusted R2              0.873               0.357                0.295
## F Statistic     175.511*** (df = 4; 44) 14.450*** (df = 4; 90) 22.947*** (df = 4; 43)
```

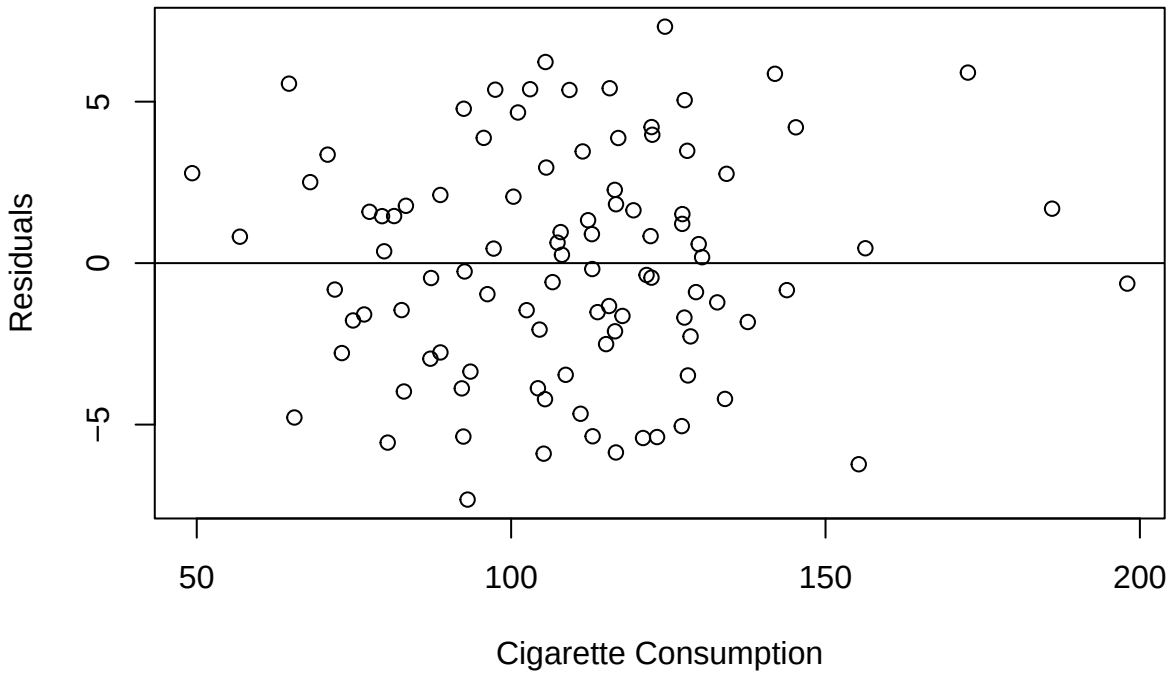
```
## ================================================================================================
## Note:                                                                *p<0.1; **p<0.05; ***p<0.01
```

From the output of the three different fixed effects models above, we can conclude that the individual effects model is better than our two ways and time effects model because it has a significantly higher R2, with a value of 0.941, while the other two models have values of 0.391 and 0.681. The R2 value of 0.941 indicates that 94.1% of the variation in the outcome variable is explained by the model.

```
# obtain regression of individual effect fixed model
fixed <- plm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price), data = data, model = "within", effect = "individual")
summary(fixed)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = packs ~ log(population) + log(income) + log(tax) +
##     log(price), data = data, effect = "individual", model = "within")
##
## Balanced Panel: n = 48, T = 2, N = 96
##
## Residuals:
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -7.3222e+00 -2.3286e+00  2.7645e-14  2.3286e+00  7.3222e+00
##
## Coefficients:
##                 Estimate Std. Error t-value  Pr(>|t|)
## log(population) -86.8005    19.3535 -4.4850 5.171e-05 ***
## log(income)      46.1164    13.6828  3.3704 0.0015713 **
## log(tax)        -10.8798     9.2993 -1.1700 0.2483161
## log(price)      -68.6026    17.1973 -3.9891 0.0002471 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    19271
## Residual Sum of Squares: 1136.6
## R-Squared:      0.94102
## Adj. R-Squared: 0.87266
## F-statistic: 175.511 on 4 and 44 DF, p-value: < 2.22e-16
```

```
# provide a residual plot for the FEM model against
# cigarette consumption
fixed.res <- resid(fixed)
plot(packs, fixed.res, ylab = "Residuals", xlab = "Cigarette Consumption")
abline(0, 0)
```



Also, from our individual effect within fixed effects model regression above, we can see that we obtain a p-value of <2.22e-16, so we can reject the null and can conclude that a within fixed effect model is appropriate for this data. In addition, we can see that when we compare the regression outputs of our lsdv and within fixed effects models, the coefficients for each variable are the same. Also, by comparing the regression outputs of the pooled OLS model and the fixed effect model, we can see that the coefficients for each variable are different. The plot of the residuals of the fixed effect model against the response variable of cigarette consumption further shows that the fixed effect model is appropriate for our data.

We can also compare the pooled model and the fixed effects model using the test below, and with the following hypotheses; Ho: Pooled model is better, Ha: Fixed effects model is better, and we reject the null hypothesis when the p-value produced from the test is less than 0.05.

```
# test the consistency of the pooled model and the within
# model
pFtest(fixed, pooled)
```

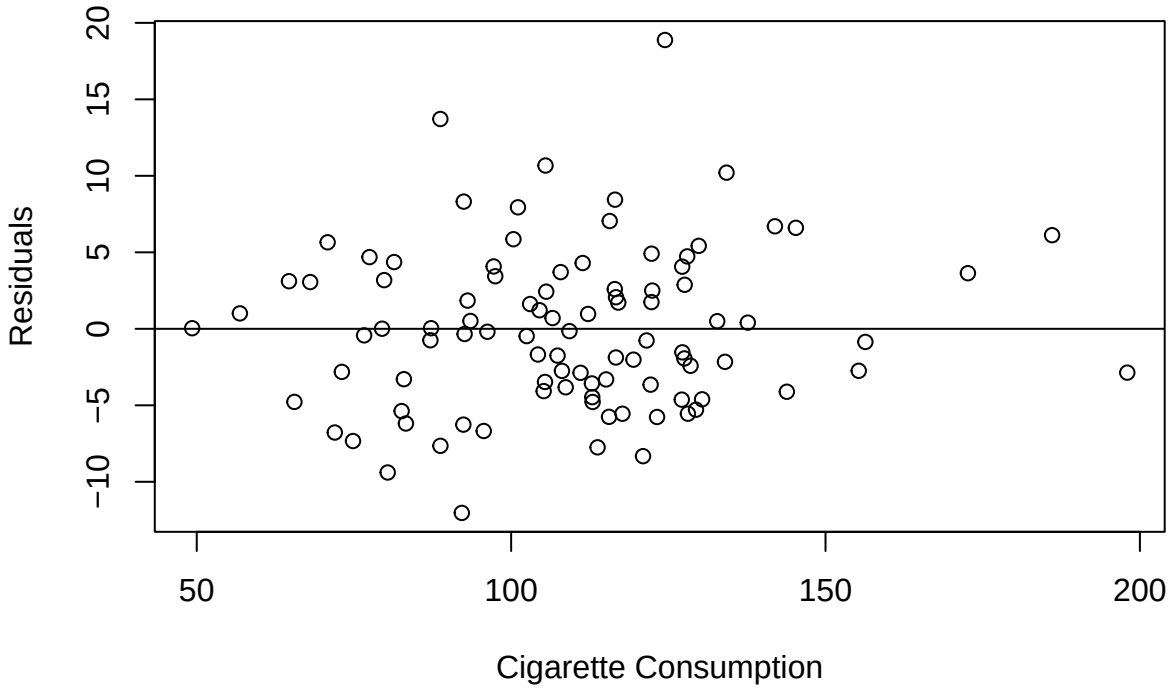
```
##
##  F test for individual effects
##
## data:  packs ~ log(population) + log(income) + log(tax) + log(price)
## F = 24.326, df1 = 47, df2 = 44, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

From the results of this test we get a p-value of <2.2e-16, and therefore reject the null and conclude that our fixed effects model is consistent and is a better fit for our data than the pooled model.

```
# fit a random effects model for the data
random <- plm(formula = packs ~ log(population) + log(income) +
    log(tax) + log(price), data = data, index = c("state", "year"),
    model = "random")
summary(random)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = packs ~ log(population) + log(income) + log(tax) +
##     log(price), data = data, model = "random", index = c("state",
##     "year"))
##
## Balanced Panel: n = 48, T = 2, N = 96
##
## Effects:
##                   var std.dev share
## idiosyncratic  25.831   5.082 0.079
## individual    303.039  17.408 0.921
## theta: 0.7978
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -12.03374  -3.69319  -0.26644   3.24571  18.87804
##
## Coefficients:
##                 Estimate Std. Error z-value  Pr(>|z|)
## (Intercept)      454.516     46.669  9.7392 < 2.2e-16 ***
## log(population)  -47.563     12.710 -3.7422 0.0001824 ***
## log(income)       42.185     11.967  3.5251 0.0004233 ***
## log(tax)         -17.880      8.170 -2.1886 0.0286294 *
## log(price)       -64.914     14.824 -4.3791 1.192e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    21083
## Residual Sum of Squares: 2608.5
## R-Squared:      0.87627
## Adj. R-Squared: 0.87084
## Chisq: 644.496 on 4 DF, p-value: < 2.22e-16
```

```r
# provide a residual plot for the REM model against
# cigarette consumption
random.res <- resid(random)
plot(packs, random.res, ylab = "Residuals", xlab = "Cigarette Consumption")
abline(0, 0)
```



The results of the regression for the random effects model tell us that the model is good, as the p-value is <2.22e-16, which is less than 0.05. We can also conclude that the individual effect of the explanatory variable "price" is statistically significant, as it is less than 0.05. In addition, the plot of the residuals shows that there might be a slight trend between the errors of the REM and the response variable of cigarette consumption, similarly to the pooled model.

We can also compare the fixed effects model and the random effects model using the Hausman test, and with the following hypotheses; Ho: REM is better, Ha: FEM is better, and we reject the null hypothesis when the p-value produced from the test is less than 0.05.

```
# conduct a Hausman test
phtest(fixed, random)
```

```
##
##  Hausman Test
##
## data:  packs ~ log(population) + log(income) + log(tax) + log(price)
## chisq = 11.79, df = 4, p-value = 0.01899
## alternative hypothesis: one model is inconsistent
```

The p-value from our Hausman test is 0.01899, which is less than 0.05, therefore we can reject the null hypothesis and conclude that the fixed effects model is better than the random effects model.

In conclusion, the various tests performed above show that the fixed effects model is better than both the pooled OLS model and the random effects model, therefore the fixed effects model is the best fit for our data.

II. Qualitative Dependent Variable Model

```r
# download qualitative dependent variable model data set
data2 <- read.csv("water_potability.csv")
ph <- data2$ph
hard <- data2$Hardness
solid <- data2$Solids
chlor <- data2$Chloramines
sulf <- data2$Sulfate
cond <- data2$Conductivity
carbon <- data2$Organic_carbon
thm <- data2$Trihalomethanes
turb <- data2$Turbidity
pot <- data2$Potability
data2 <- data.frame(ph, hard, solid, chlor, sulf, cond, carbon,
    thm, turb, pot)

# remove rows with NAs from the data set
data2 <- na.omit(data2)
```

(1) This data set is from Kaggle, and gives data on the quality of 3272 samples of drinking water, depending on a variety of explanatory variables. There are 9 explanatory variables; "ph", which tells us the acid-base balance of water, where a ph value lower than 7 indicates that the water is more acidic, and a value greater than 7 is more basic; "hard", which measures the capacity of water needed to precipitate soap caused by calcium and magnesium; "solid", which measures the total dissolved mineral solids in the water, where a high value indicates the water is highly mineralized; "chlor", which measures the amount of chloramines and chlorine in the water; "sulf", which measures the amount of sulfates present in the water; "cond", which measures the electrical conductivity of the water; "carbon", which is the total organic carbon in the water; "thm", which is the level of trihalomethanes in the water, dependent on the level of organic material in the water, amount of chlorine required to treat the water, and the temperature of the water being treated; "turb", which measures the turbidity of the water, which depends on the quantity of solid matter in the suspended state of the water. These explanatory variables help to explain the binary dependent variable "pot", which is the potability of the water, and indicates if the water is safe for consumption, where a value of "1" means the water is potable/drinkable, and "0" means the water is not potable/drinkable.

(2) Provide a descriptive analysis of all of the variables, including histograms, fitted distributions, correlation plots, box plots, scatterplots and statistical summaries.

```r
# provide five-number for each variable
summary(data2)
```

```
##        ph              hard            solid             chlor
##  Min.   : 0.2275   Min.   : 73.49   Min.   :  320.9   Min.   : 1.391
##  1st Qu.: 6.0897   1st Qu.:176.74   1st Qu.:15615.7   1st Qu.: 6.139
##  Median : 7.0273   Median :197.19   Median :20933.5   Median : 7.144
##  Mean   : 7.0860   Mean   :195.97   Mean   :21917.4   Mean   : 7.134
##  3rd Qu.: 8.0530   3rd Qu.:216.44   3rd Qu.:27182.6   3rd Qu.: 8.110
##  Max.   :14.0000   Max.   :317.34   Max.   :56488.7   Max.   :13.127
##       sulf            cond           carbon           thm
##  Min.   :129.0   Min.   :201.6   Min.   : 2.20   Min.   :  8.577
##  1st Qu.:307.6   1st Qu.:366.7   1st Qu.:12.12   1st Qu.: 55.953
```

```
##  Median :332.2   Median :423.5   Median :14.32   Median : 66.542
##  Mean   :333.2   Mean   :426.5   Mean   :14.36   Mean   : 66.401
##  3rd Qu.:359.3   3rd Qu.:482.4   3rd Qu.:16.68   3rd Qu.: 77.292
##  Max.   :481.0   Max.   :753.3   Max.   :27.01   Max.   :124.000
##       turb            pot
##  Min.   :1.450   Min.   :0.0000
##  1st Qu.:3.443   1st Qu.:0.0000
##  Median :3.968   Median :0.0000
##  Mean   :3.970   Mean   :0.4033
##  3rd Qu.:4.514   3rd Qu.:1.0000
##  Max.   :6.495   Max.   :1.0000
```

```
# find the percentage of potable vs. not potable water in
# the data
length(data2$pot[data2$pot == 0])
```

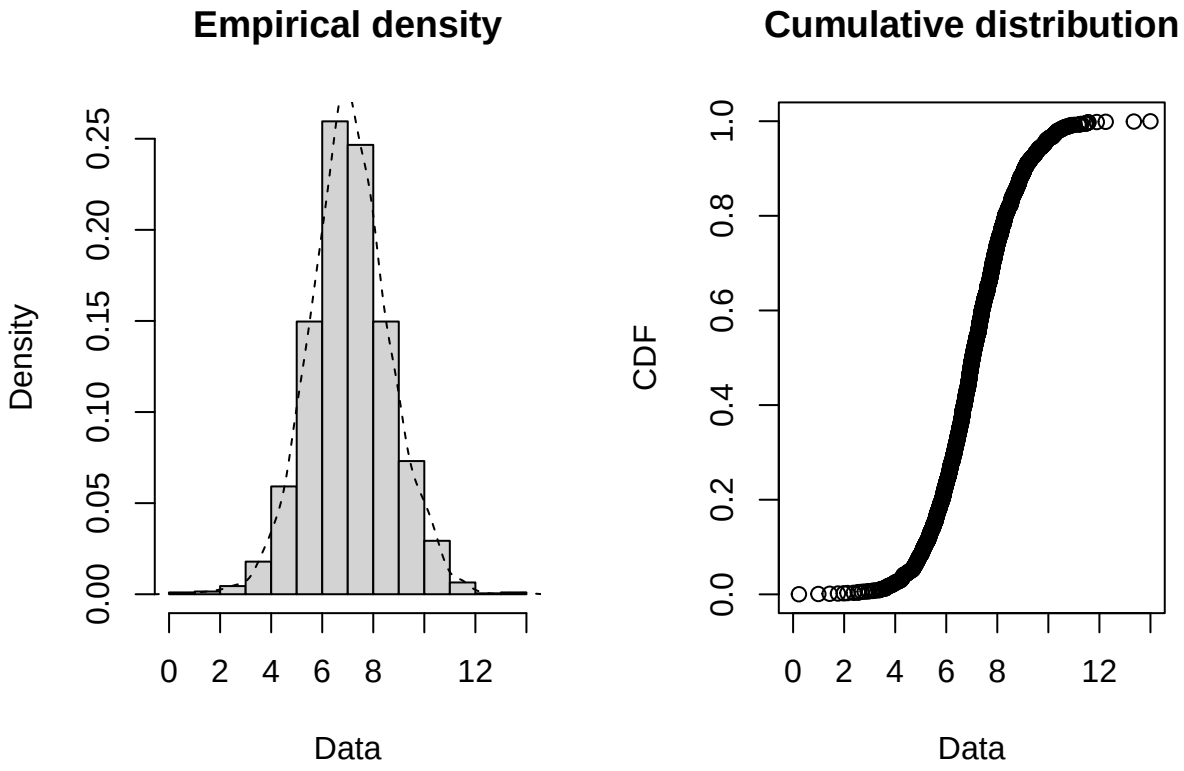
```
## [1] 1200
```

```
length(data2$pot[data2$pot == 1])
```

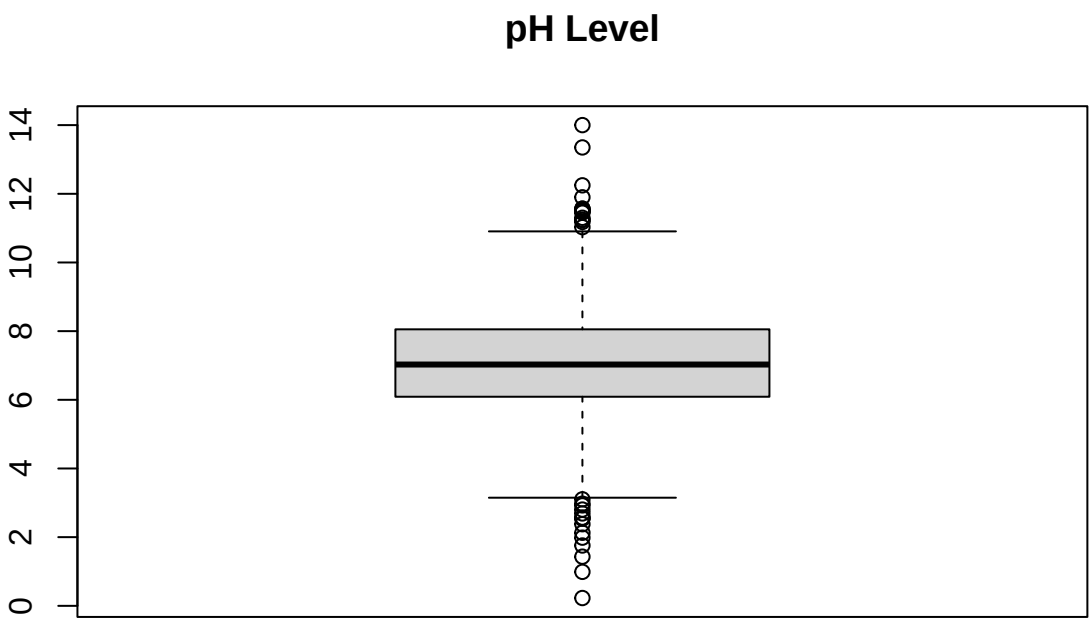
```
## [1] 811
```

There are a total of 2011 observations in the data set, and 1200 of these observations have a "pot" value of 0 and 811 have a "pot" value of 1, i.e. approximately 59.67% of the water samples taken are not potable/safe to drink, and the other 40.33% of the water samples are potable/drinkable.

```
# provide a histogram and box plot for the ph level
plotdist(data2$ph, histo = TRUE, demp = TRUE)
```

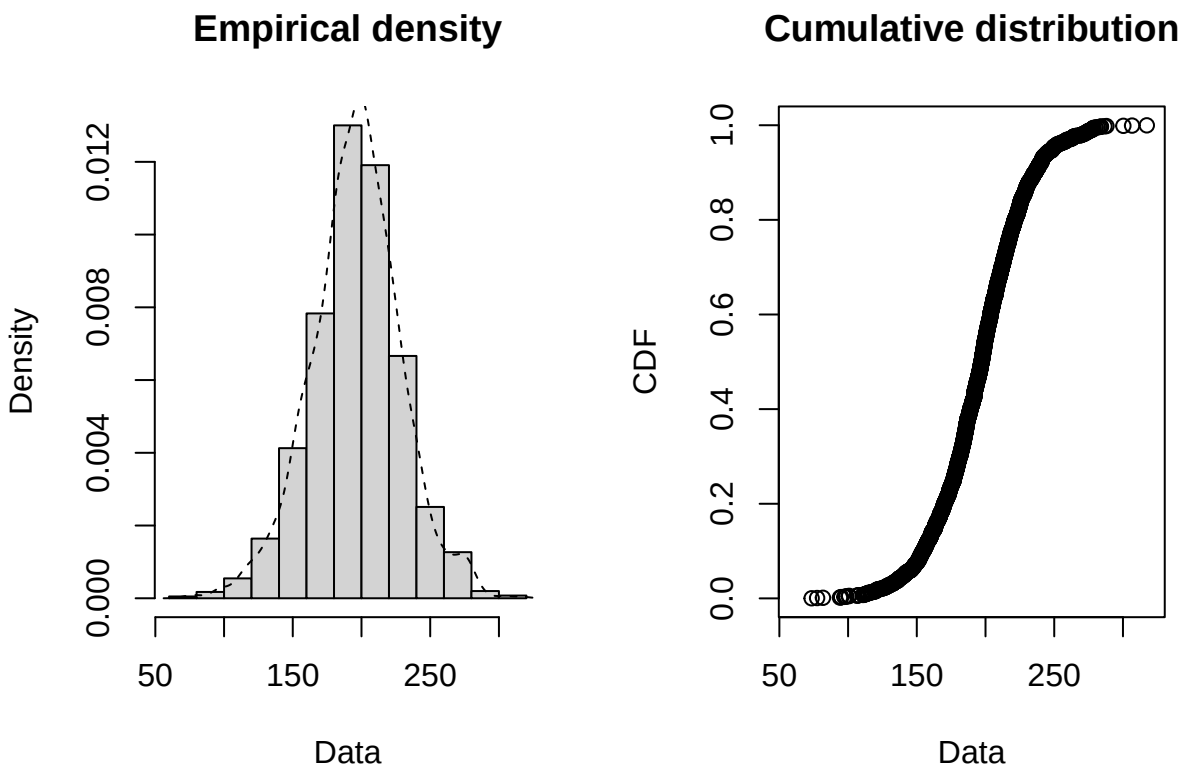


```
boxplot(data2$ph, main = "pH Level")
```

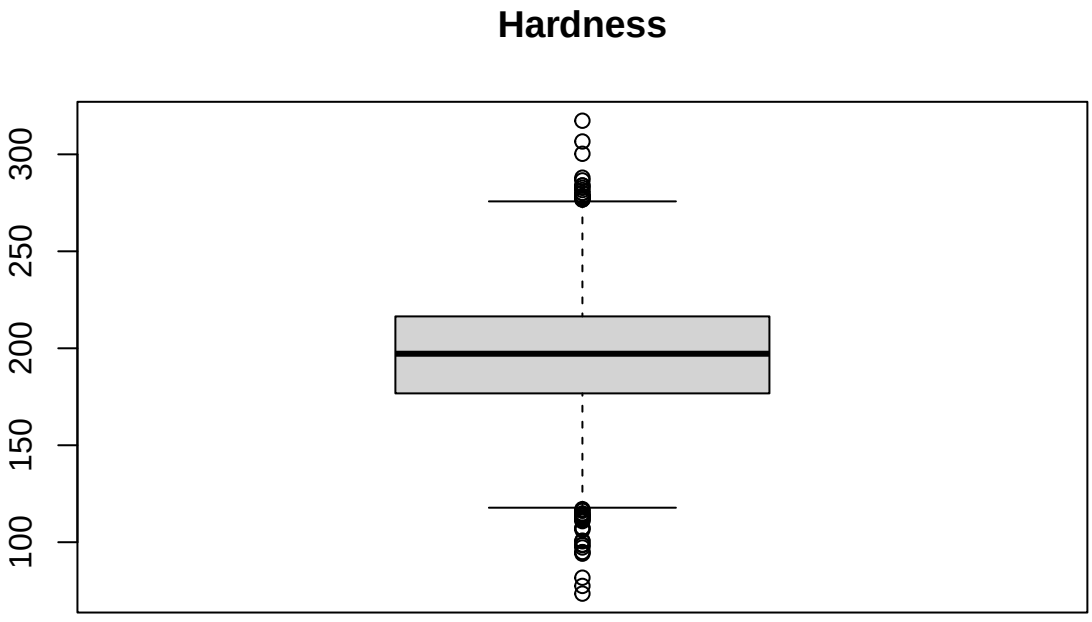


The histogram for pH level show that it has a relatively Normal distribution with a mode of approximately 7, with the mean pH level being very close to neutral, with a few values that are either very acidic or very basic. In addition, from the box plot for pH level, we can see that there are outliers on both the high end and low end for pH level.

```
# provide a histogram and box plot for the hardness of the
# water
plotdist(data2$hard, histo = TRUE, demp = TRUE)
```

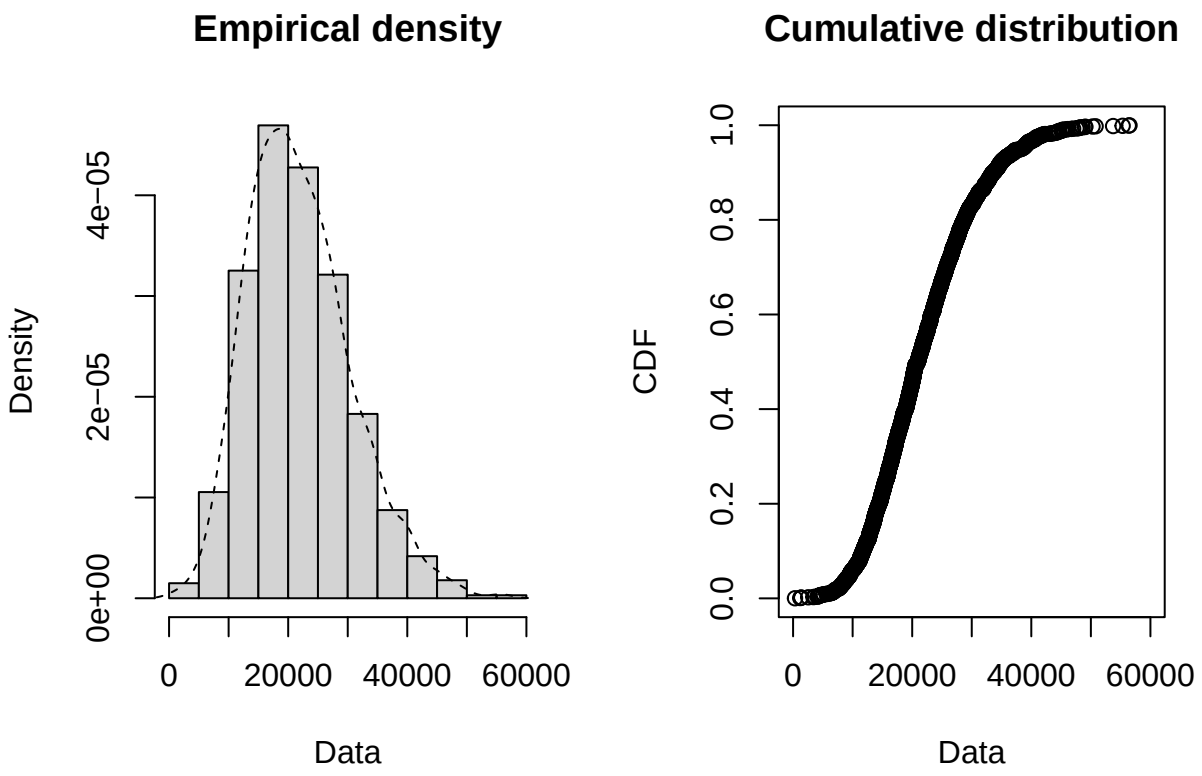



```
boxplot(data2$hard, main = "Hardness")
```

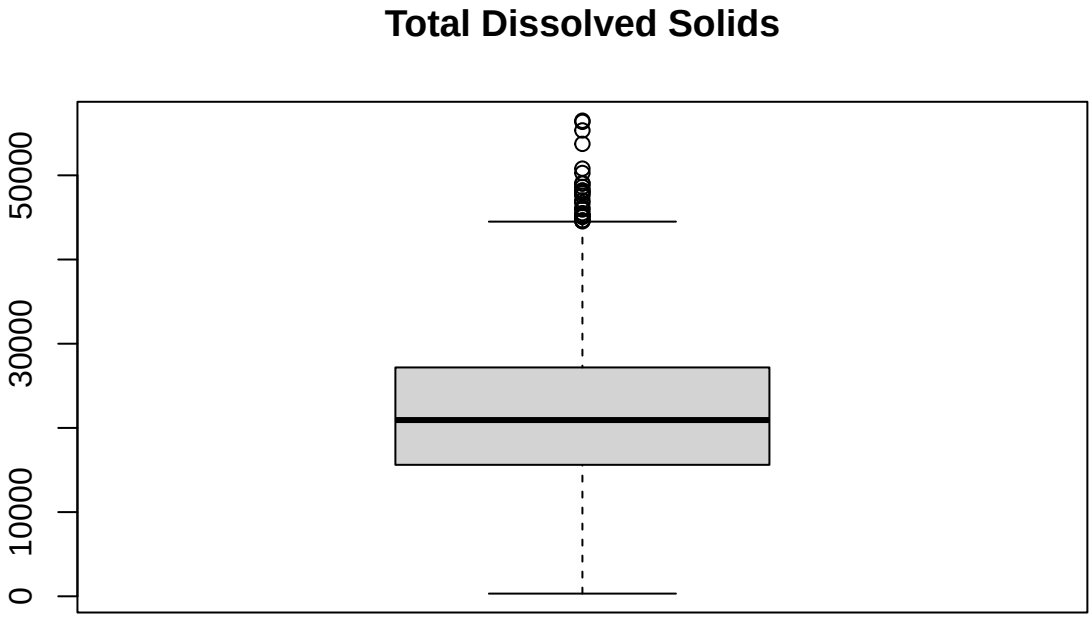


The histogram for the hardness of the water shows a relatively Normal distribution, but very slightly skewed left, with most hardness values being around approximately 200. Furthermore, the box plot also shows an even distribution of the hardness level values, and also that there are both high and low outliers.

```
# provide a histogram and box plot for the total dissolved
# solids in the water
plotdist(data2$solid, histo = TRUE, demp = TRUE)
```

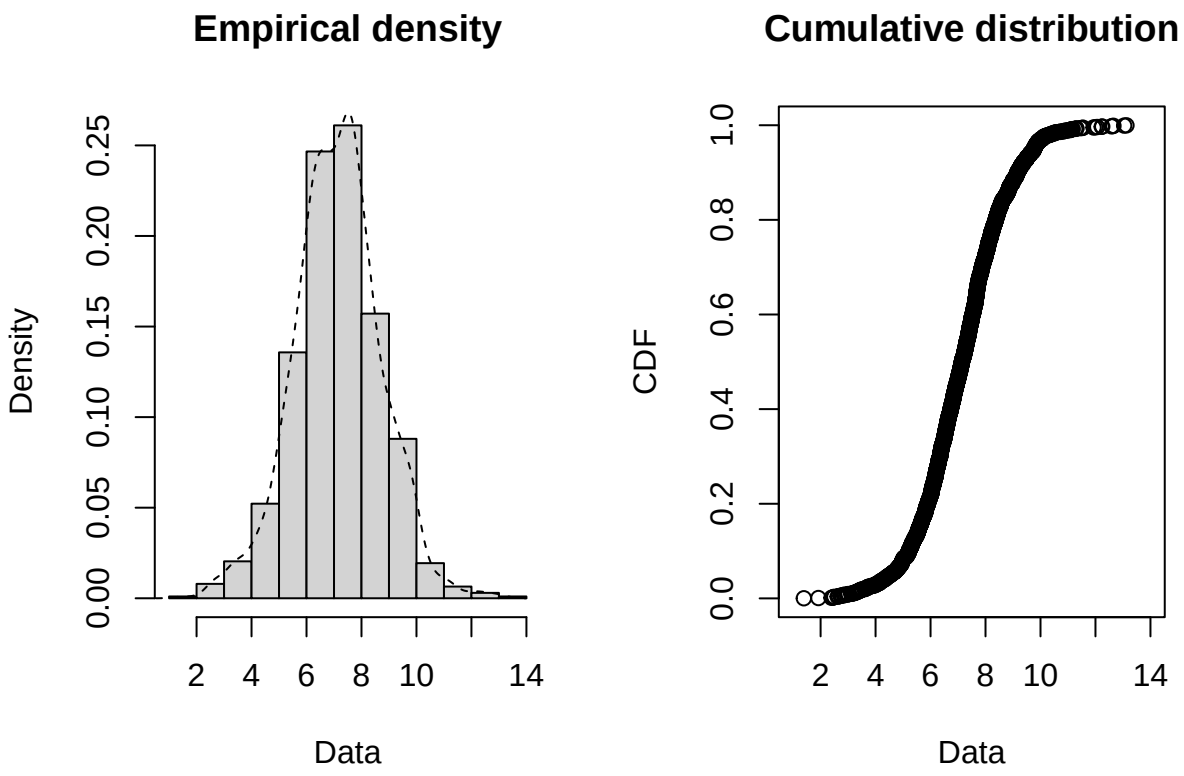



```
boxplot(data2$solid, main = "Total Dissolved Solids")
```

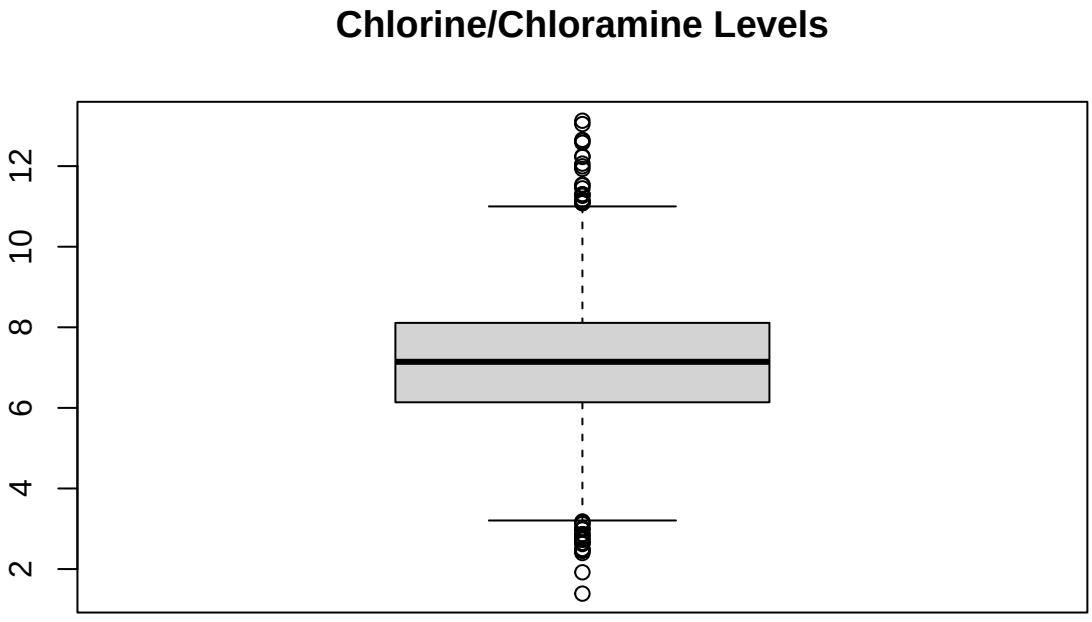


The histogram for total dissolved solid levels is skewed right, with a mode of approximately 15,000 mg/L. The box plot for total dissolved solids also shows that there are outliers on the higher side, and if these outliers were removed, the distribution would likely resemble a Normal distribution but with a lower maximum.

```
# provide a histogram and box plot for chlorine and
# chloramine levels
plotdist(data2$chlor, histo = TRUE, demp = TRUE)
```

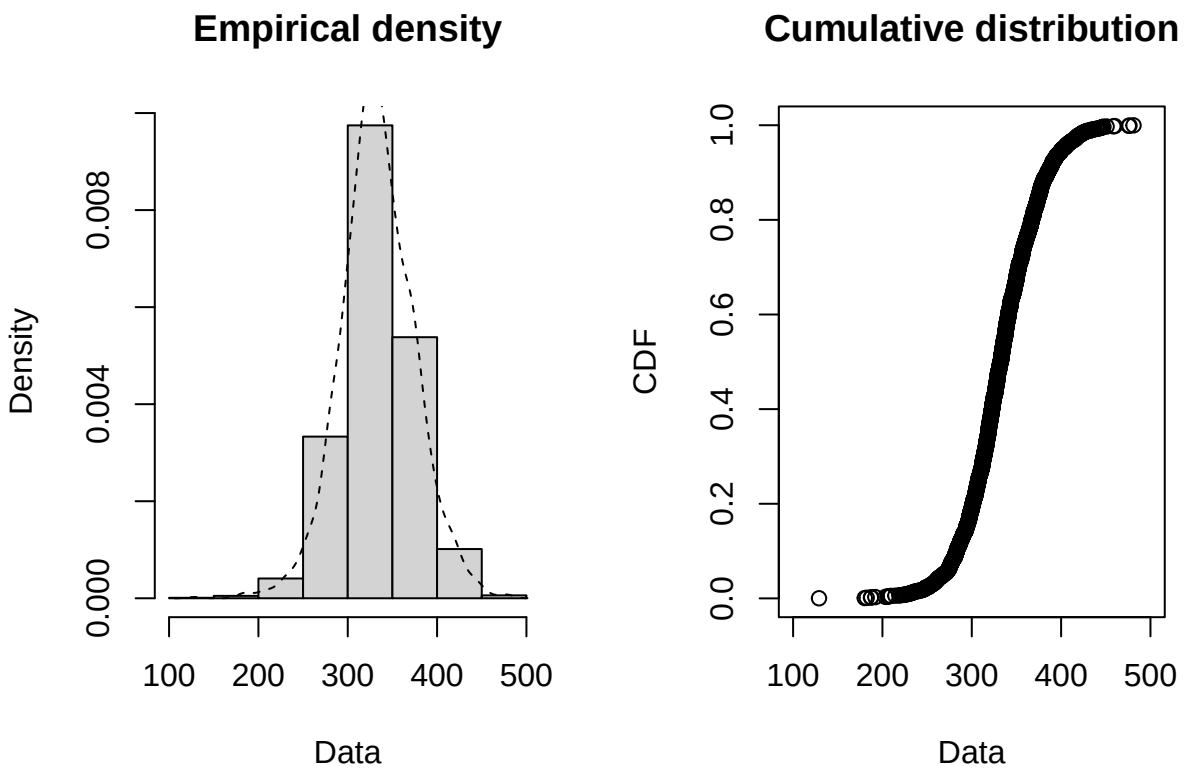



```
boxplot(data2$chlor, main = "Chlorine/Chloramine Levels")
```

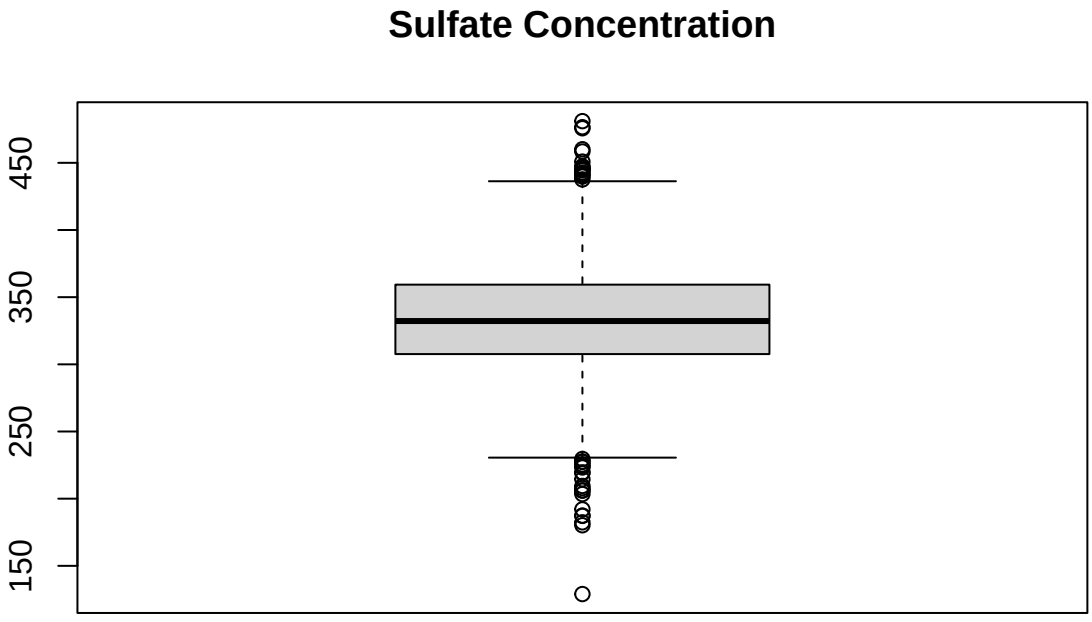


The histogram for chlorine/chloramine levels is Normally distributed, with a mean and mode of approximately 7-8. In addition, the box plot for chlorine/chloramine levels shows both high and low outliers, but if these outliers were removed, the distribution would still be Normal, just over a shorter range.

```
# provide a histogram and box plot for sulfate levels
plotdist(data2$sulf, histo = TRUE, demp = TRUE)
```

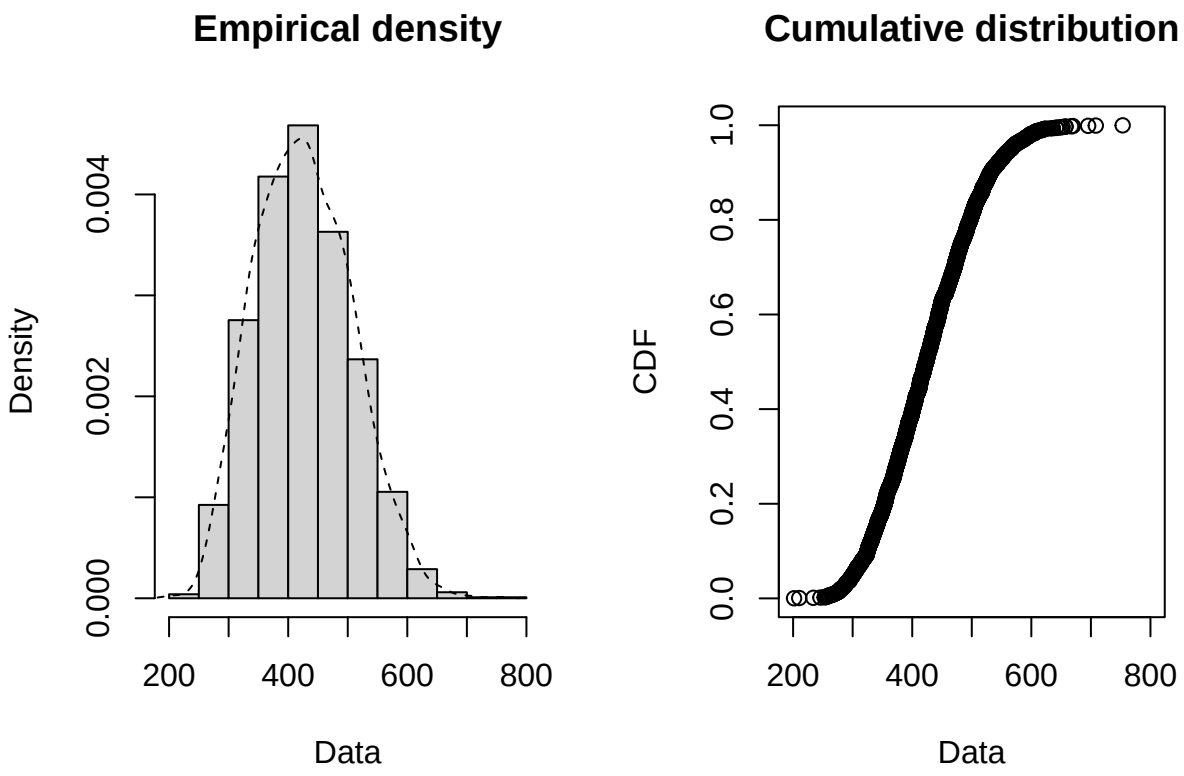



```
boxplot(data2$sulf, main = "Sulfate Concentration")
```

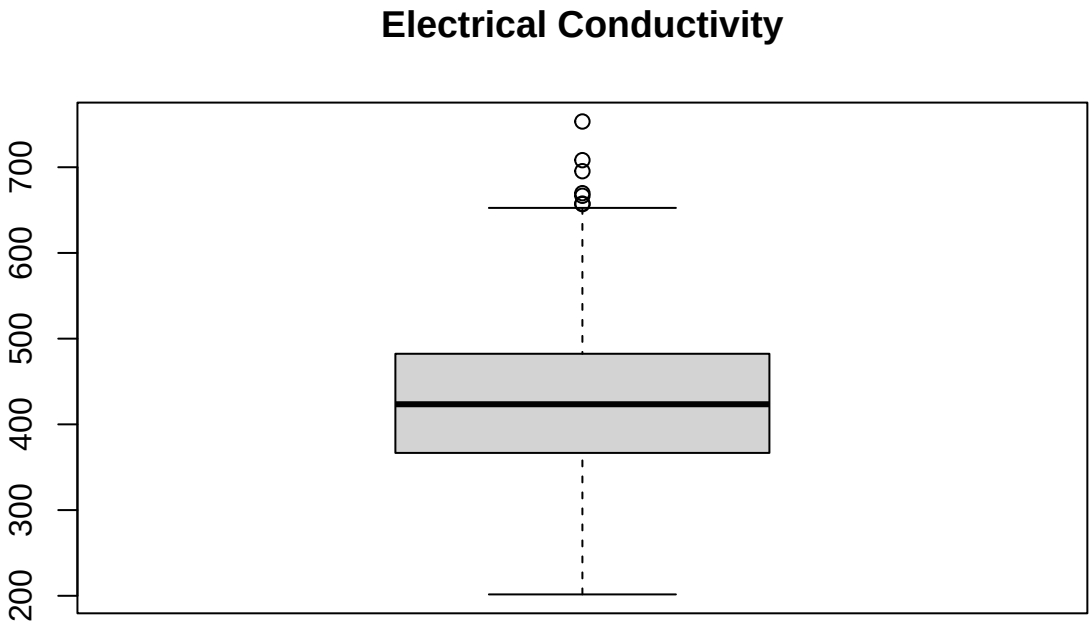


The histogram for sulfate concentration shows a relatively Normal distribution, with most of the values falling around 300-350 mg/L. Furthermore, the box plot for sulfate concentration shows both high and low outliers, but if these outliers were removed, the distribution for sulfate concentration would likely still be Normal, just over a shorter range.

```
# provide a histogram and box plot for the electrical
# conductivity of the water
plotdist(data2$cond, histo = TRUE, demp = TRUE)
```

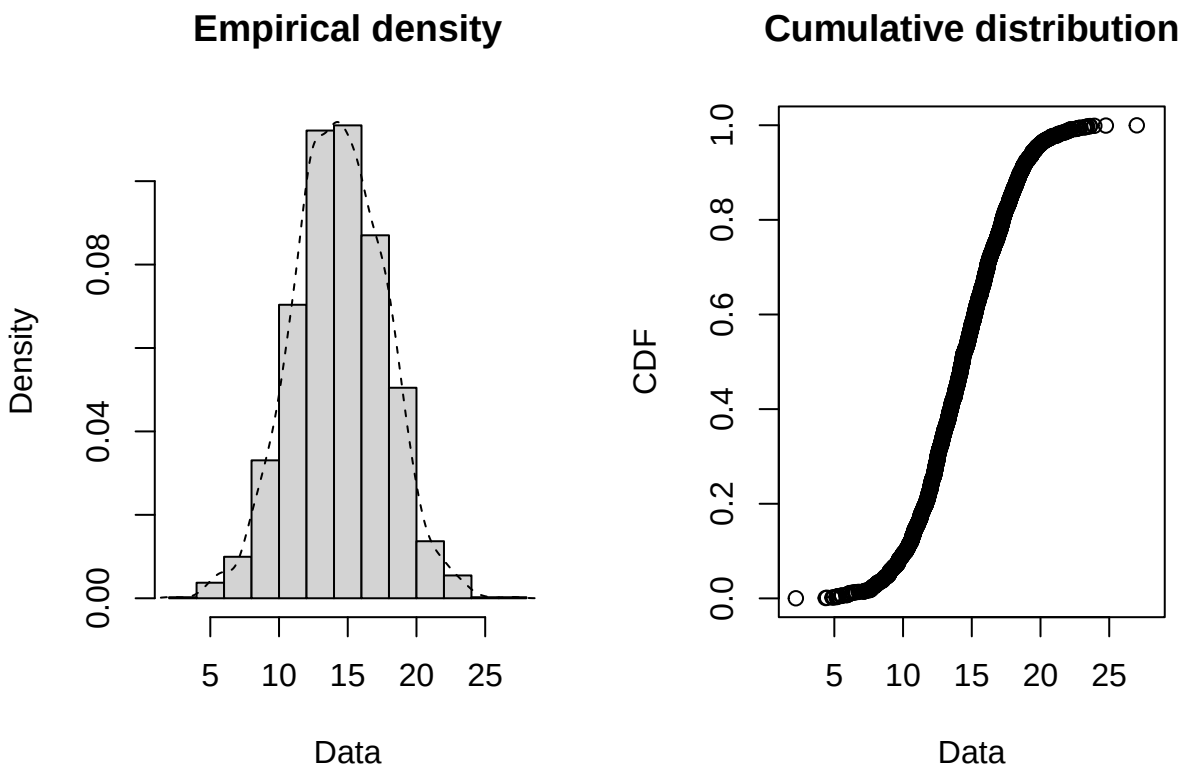



```
boxplot(data2$cond, main = "Electrical Conductivity")
```

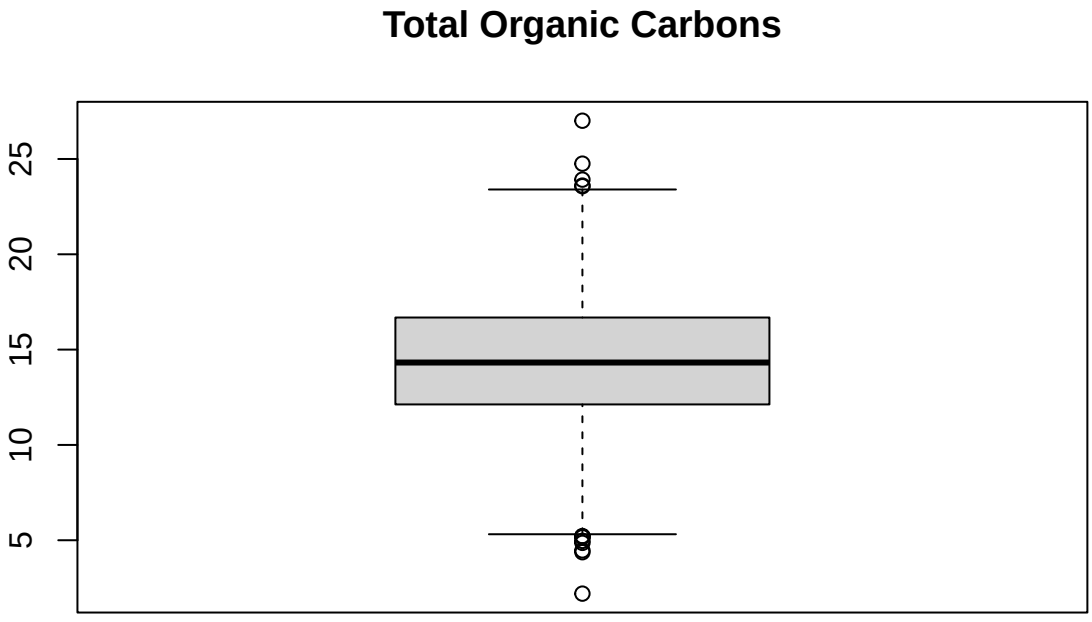


The histogram for electrical conductivity shows that the distribution is slightly skewed right, but the box plot shows that if the outliers, which are all on the high side, were removed, the distribution would become Normal.

```
# provide a histogram and box plot for the total organic
# carbon levels
plotdist(data2$carbon, histo = TRUE, demp = TRUE)
```

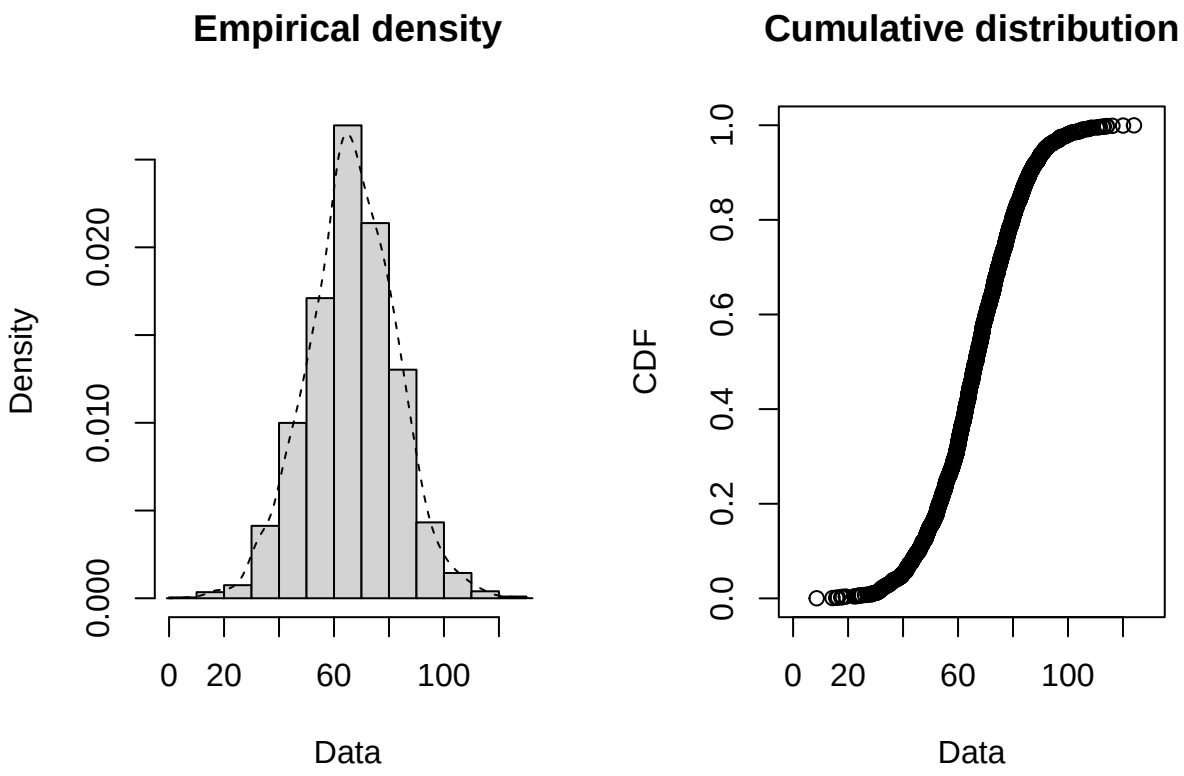



```
boxplot(data2$carbon, main = "Total Organic Carbons")
```

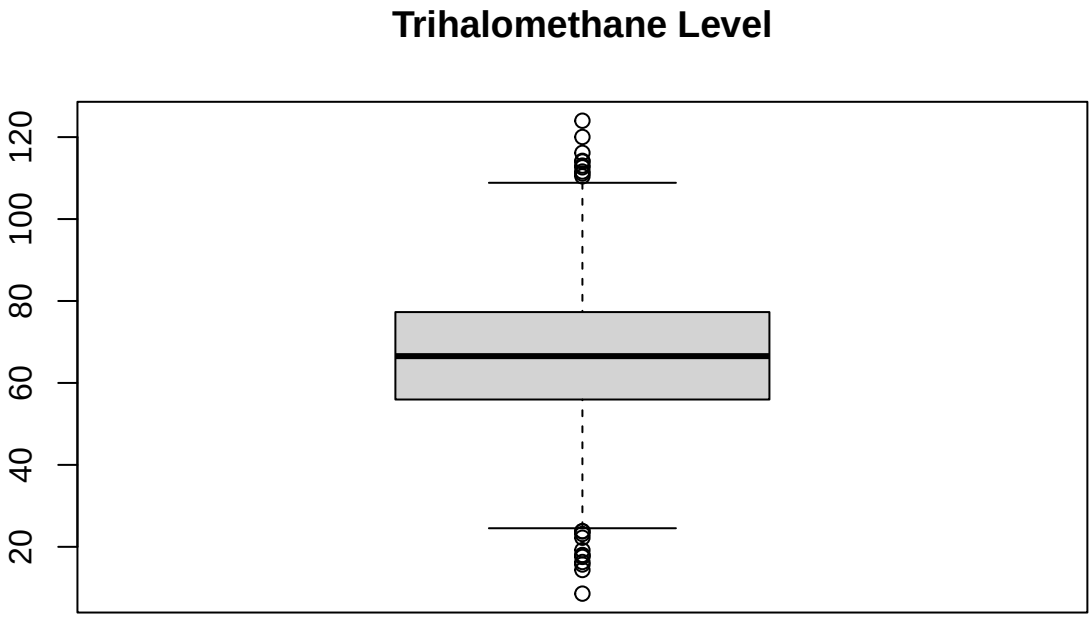


The histogram and box plot for total organic carbons in the water show that there is a Normal distribution, with outliers on both the high and low end, but if these outliers were removed, the distribution would likely still be Normal, with most of the values falling around 15 mg/L.

```
# provide a histogram and box plot for the trihalomethane
# levels
plotdist(data2$thm, histo = TRUE, demp = TRUE)
```

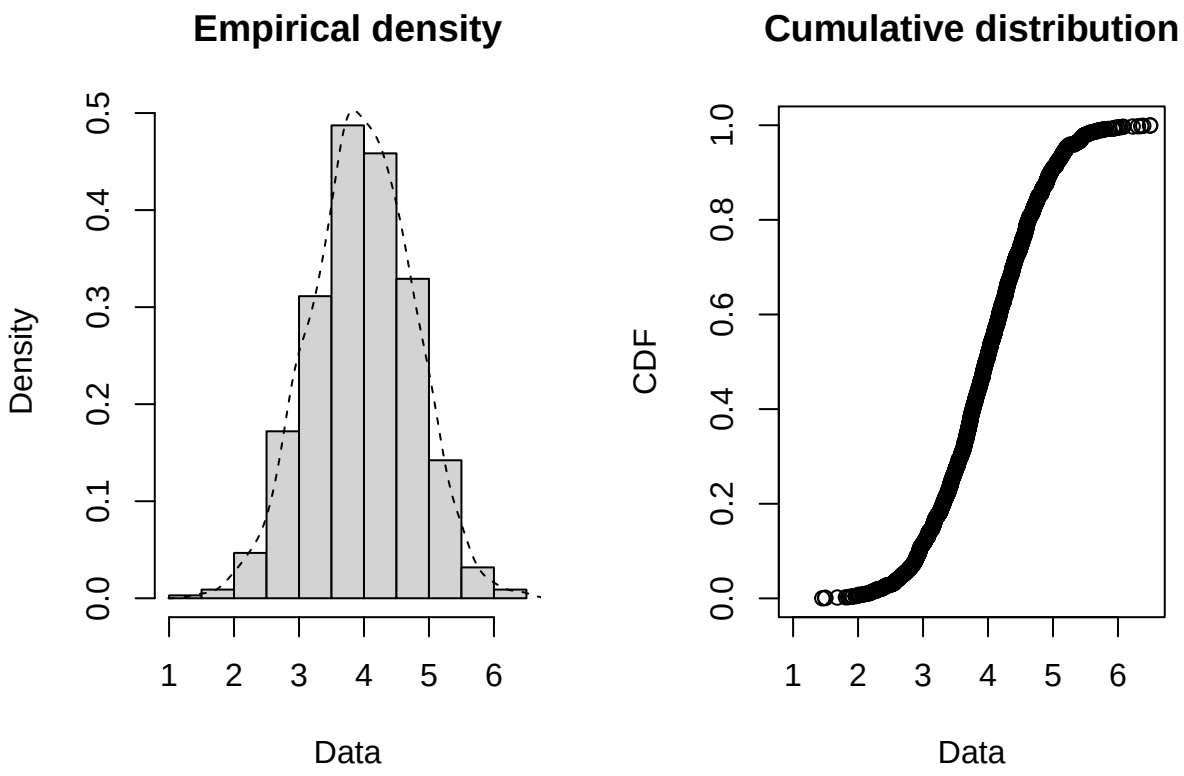



```
boxplot(data2$thm, main = "Trihalomethane Level")
```

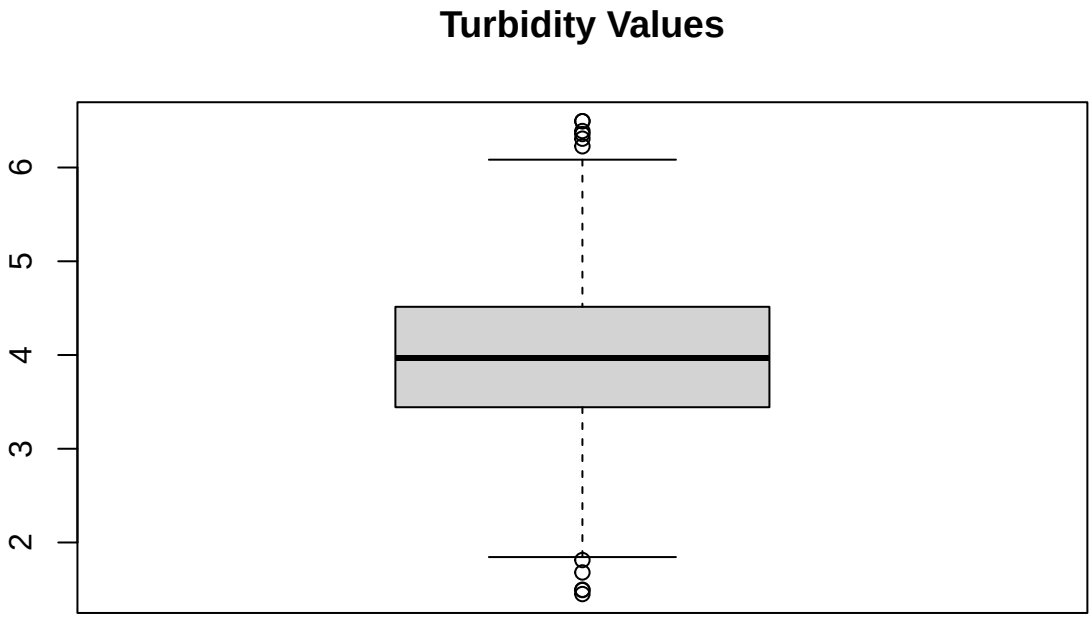


The histogram and box plot for trihalomethane levels in the water show that there is a Normal distribution, with outliers on both the high and low end, but if these outliers were removed, the distribution would likely still be Normal, with most of the values falling around 60-70 ppm.

```
# provide a histogram and box plot for the turbidity of the
# water
plotdist(data2$turb, histo = TRUE, demp = TRUE)
```

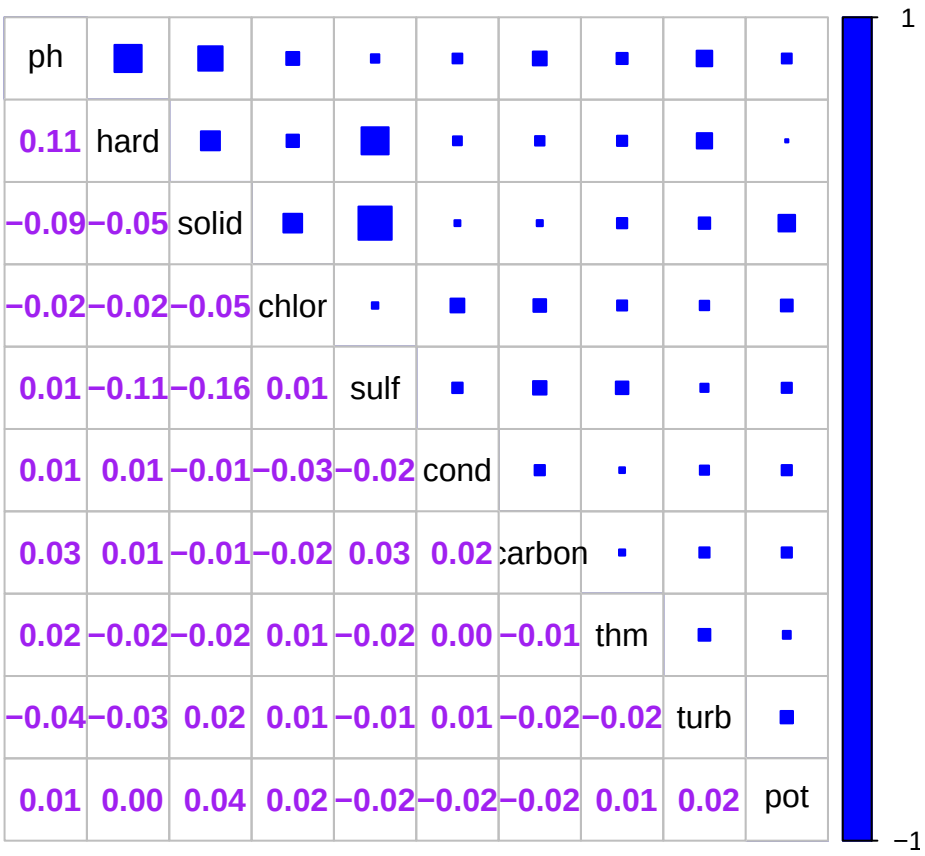



```
boxplot(data2$turb, main = "Turbidity Values")
```



The histogram and box plot for turbidity levels in the water show that the distribution is relatively Normal, but very slightly skewed left, with outliers on both the high and low end, but if these outliers were removed, the distribution would likely still be relatively Normal, with most of the values falling around 4 NTU.

```
# provide a correlation plot of the variables
corrplot.mixed(cor(data2), upper = "square", lower = "number",
    addgrid.col = "black", tl.col = "black", lower.col = "purple",
    upper.col = "blue")
```



From the correlation plot, we can see that all of the correlation values between each of the variables are very low, or not close to 1 or -1, indicating that the variables are not highly correlated, and that they are all very important predictors for our model.

(3) Fit a linear probability model, a probability unit (probit) model, and a logistic (logit) model to conclude which one is preferred for the data. Include other evidence (plots, statistical diagnostics, etc.) to support the conclusion.

```
# fit a basic OLS model
lpm <- lm(pot ~ ph + hard + solid + chlor + sulf + cond + carbon +
    thm + turb, data = data2)
summary(lpm)
```

```
##
## Call:
## lm(formula = pot ~ ph + hard + solid + chlor + sulf + cond +
##     carbon + thm + turb, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5031 -0.4068 -0.3745  0.5870  0.7012
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.875e-01  1.797e-01   1.600    0.110
```

```
## ph           6.283e-03  7.035e-03   0.893    0.372
## hard        -8.730e-06  3.406e-04  -0.026    0.980
## solid        2.371e-06  1.294e-06   1.832    0.067 .
## chlor        6.897e-03  6.929e-03   0.995    0.320
## sulf        -9.925e-05  2.715e-04  -0.366    0.715
## cond        -9.216e-05  1.358e-04  -0.679    0.497
## carbon      -2.142e-03  3.297e-03  -0.650    0.516
## thm          2.882e-04  6.818e-04   0.423    0.673
## turb         1.406e-02  1.406e-02   1.000    0.317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4909 on 2001 degrees of freedom
## Multiple R-squared:  0.003638,   Adjusted R-squared:  -0.0008434
## F-statistic: 0.8118 on 9 and 2001 DF,  p-value: 0.6053
```

```
kable(tidy(lpm), digits = 4, align = "c", caption = "Linear Probability Model")
```

Table 1: Linear Probability Model

term	estimate	std.error	statistic	p.value
(Intercept)	0.2875	0.1797	1.5996	0.1098
ph	0.0063	0.0070	0.8931	0.3719
hard	0.0000	0.0003	-0.0256	0.9796
solid	0.0000	0.0000	1.8325	0.0670
chlor	0.0069	0.0069	0.9954	0.3197
sulf	-0.0001	0.0003	-0.3655	0.7148
cond	-0.0001	0.0001	-0.6788	0.4974
carbon	-0.0021	0.0033	-0.6495	0.5161
thm	0.0003	0.0007	0.4227	0.6726
turb	0.0141	0.0141	1.0002	0.3174

```
# fit a linear probability model using heteroskedastic
# consistent errors
cov1 <- hccm(lpm, type = "hc1")
hcerr <- coeftest(lpm, vcov. = cov1)
hcerr
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept)  2.8749e-01  1.8093e-01  1.5890  0.11222
## ph           6.2831e-03  6.9181e-03  0.9082  0.36387
## hard        -8.7298e-06  3.5102e-04 -0.0249  0.98016
## solid        2.3711e-06  1.2832e-06  1.8477  0.06479 .
## chlor        6.8966e-03  7.1079e-03  0.9703  0.33203
## sulf        -9.9248e-05  2.8216e-04 -0.3517  0.72506
## cond        -9.2163e-05  1.3665e-04 -0.6745  0.50009
## carbon      -2.1417e-03  3.2605e-03 -0.6569  0.51134
## thm          2.8820e-04  6.8531e-04  0.4205  0.67414
## turb         1.4058e-02  1.4062e-02  0.9997  0.31757
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kable(tidy(hcerr), digits = 4, align = "c", caption = "Linear Probability Model with HC Errors")
```

Table 2: Linear Probability Model with HC Errors

term	estimate	std.error	statistic	p.value
(Intercept)	0.2875	0.1809	1.5890	0.1122
ph	0.0063	0.0069	0.9082	0.3639
hard	0.0000	0.0004	-0.0249	0.9802
solid	0.0000	0.0000	1.8477	0.0648
chlor	0.0069	0.0071	0.9703	0.3320
sulf	-0.0001	0.0003	-0.3517	0.7251
cond	-0.0001	0.0001	-0.6745	0.5001
carbon	-0.0021	0.0033	-0.6569	0.5113
thm	0.0003	0.0007	0.4205	0.6741
turb	0.0141	0.0141	0.9997	0.3176

```
# fit a probit model
probit <- glm(pot ~ ph + hard + solid + chlor + sulf + cond +
    carbon + thm + turb, family = binomial(link = "probit"),
    data = data2)
summary(probit)
```

```
##
## Call:
## glm(formula = pot ~ ph + hard + solid + chlor + sulf + cond +
##     carbon + thm + turb, family = binomial(link = "probit"),
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1843  -1.0217  -0.9693   1.3313   1.5405
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.538e-01  4.648e-01  -1.192   0.2335
## ph           1.555e-02  1.818e-02   0.856   0.3922
## hard        -3.559e-07  8.800e-04   0.000   0.9997
## solid        6.111e-06  3.338e-06   1.831   0.0671 .
## chlor        1.759e-02  1.791e-02   0.982   0.3259
## sulf        -2.180e-04  7.014e-04  -0.311   0.7560
## cond        -2.383e-04  3.511e-04  -0.679   0.4973
## carbon      -5.560e-03  8.521e-03  -0.653   0.5140
## thm          7.342e-04  1.762e-03   0.417   0.6770
## turb         3.632e-02  3.633e-02   1.000   0.3174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 2712.1  on 2010  degrees of freedom
## Residual deviance: 2704.9  on 2001  degrees of freedom
## AIC: 2724.9
##
## Number of Fisher Scoring iterations: 4
```

```
kable(tidy(probit), digits = 4, align = "c", caption = "Probit Model")
```

Table 3: Probit Model

term	estimate	std.error	statistic	p.value
(Intercept)	-0.5538	0.4648	-1.1915	0.2335
ph	0.0156	0.0182	0.8556	0.3922
hard	0.0000	0.0009	-0.0004	0.9997
solid	0.0000	0.0000	1.8308	0.0671
chlor	0.0176	0.0179	0.9823	0.3259
sulf	-0.0002	0.0007	-0.3108	0.7560
cond	-0.0002	0.0004	-0.6787	0.4973
carbon	-0.0056	0.0085	-0.6526	0.5140
thm	0.0007	0.0018	0.4166	0.6770
turb	0.0363	0.0363	0.9997	0.3174

```
# fit a logit model
logit <- glm(pot ~ ph + hard + solid + chlor + sulf + cond +
    carbon + thm + turb, family = binomial(link = "logit"), data = data2)
summary(logit)
```

```
##
## Call:
## glm(formula = pot ~ ph + hard + solid + chlor + sulf + cond +
##     carbon + thm + turb, family = binomial(link = "logit"), data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1868  -1.0214  -0.9687   1.3304   1.5431
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.758e-01  7.487e-01  -1.170   0.2421
## ph           2.615e-02  2.926e-02   0.894   0.3716
## hard        -3.596e-05  1.417e-03  -0.025   0.9798
## solid        9.842e-06  5.368e-06   1.833   0.0668 .
## chlor        2.876e-02  2.884e-02   0.997   0.3187
## sulf        -4.111e-04  1.129e-03  -0.364   0.7158
## cond        -3.853e-04  5.657e-04  -0.681   0.4958
## carbon      -8.912e-03  1.372e-02  -0.650   0.5160
## thm          1.200e-03  2.838e-03   0.423   0.6725
## turb         5.873e-02  5.851e-02   1.004   0.3154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2712.1  on 2010  degrees of freedom
## Residual deviance: 2704.8  on 2001  degrees of freedom
## AIC: 2724.8
##
## Number of Fisher Scoring iterations: 4
```

```
kable(tidy(logit), digits = 4, align = "c", caption = "Logit Model")
```

Table 4: Logit Model

term	estimate	std.error	statistic	p.value
(Intercept)	-0.8758	0.7487	-1.1697	0.2421
ph	0.0261	0.0293	0.8935	0.3716
hard	0.0000	0.0014	-0.0254	0.9798
solid	0.0000	0.0000	1.8333	0.0668
chlor	0.0288	0.0288	0.9971	0.3187
sulf	-0.0004	0.0011	-0.3641	0.7158
cond	-0.0004	0.0006	-0.6812	0.4958
carbon	-0.0089	0.0137	-0.6496	0.5160
thm	0.0012	0.0028	0.4227	0.6725
turb	0.0587	0.0585	1.0039	0.3154

```
# compare the three models
stargazer(hcerr, probit, logit, header = FALSE, title = "Three Binary Choice Models",
    type = "text", keep.stat = "n", digits = 4, single.row = FALSE,
    intercept.bottom = FALSE, model.names = FALSE, column.labels = c("LPM",
        "probit", "logit"), omit.table.layout = "n")
```

```
##
## Three Binary Choice Models
## ============================================
##                    Dependent variable:
##               ------------------------------
##                              pot
##                  LPM       probit     logit
##                  (1)        (2)        (3)
## --------------------------------------------
## Constant       0.2875    -0.5538    -0.8758
##               (0.1809)   (0.4648)  (0.7487)
##
## ph             0.0063     0.0156     0.0261
##               (0.0069)   (0.0182)  (0.0293)
##
## hard          -0.00001  -0.000000  -0.00004
##               (0.0004)   (0.0009)  (0.0014)
##
## solid        0.000002*   0.00001*  0.00001*
##              (0.000001) (0.000003) (0.00001)
##
```

```
## chlor          0.0069     0.0176    0.0288
##               (0.0071)   (0.0179)  (0.0288)
##
## sulf          -0.0001    -0.0002    -0.0004
##               (0.0003)   (0.0007)  (0.0011)
##
## cond          -0.0001    -0.0002    -0.0004
##               (0.0001)   (0.0004)  (0.0006)
##
## carbon        -0.0021    -0.0056    -0.0089
##               (0.0033)   (0.0085)  (0.0137)
##
## thm            0.0003     0.0007    0.0012
##               (0.0007)   (0.0018)  (0.0028)
##
## turb           0.0141     0.0363    0.0587
##               (0.0141)   (0.0363)  (0.0585)
##
## ---------------------------------------------
## Observations               2,011      2,011
## =============================================
```

The regression output for the three binary choice models shows us that the coefficient estimates for all of the explanatory variables are extremely low and very close to zero, and the individual effects for each of the estimation parameters are insignificant in all models, with the exception of the variable "solid" that measures the quantity of total dissolved solids in the water.

We can further compare the three models using a hypothetical predicted probability of 0.5.

```
# comparing the three models using a hypothetical threshold
# of predicted probability of 0.5
(tab <- table(predict(lpm) > 0.5, data2$pot))
```

```
##
##            0    1
##   FALSE 1199  808
##   TRUE     1    3
```

```
c(tab[1, 1]/sum(tab[, 1]), tab[2, 2]/sum(tab[, 2]))
```

```
## [1] 0.999166667 0.003699137
```

```
(tab <- table(predict(probit, type = "response") > 0.5, data2$pot))
```

```
##
##            0    1
##   FALSE 1198  808
##   TRUE     2    3
```

```
c(tab[1, 1]/sum(tab[, 1]), tab[2, 2]/sum(tab[, 2]))
```

```
## [1] 0.998333333 0.003699137
```

```
(tab <- table(predict(logit, type = "response") > 0.5, data2$pot))
```

```
##
##            0    1
##   FALSE 1198  807
##   TRUE     2    4
```

```
c(tab[1, 1]/sum(tab[, 1]), tab[2, 2]/sum(tab[, 2]))
```

```
## [1] 0.998333333 0.004932182
```

For the results above, the first number gives the share of not potable water samples predicted correctly (about 99.9%), and the second number gives the share of potable water samples predicted correctly (about 0.4%). We can see that the LPM model is slightly better at predicting not potable water samples, and the probit and logit models are equally as good, and that the logit model is slightly better at predicting potable water samples, with the LPM and probit models being equally as good. From these results we can conclude that the logit model is the best fit for our data.

```
# find average marginal effect of each explanatory variable
# for all 3 models
margins(lpm)
```

```
## Average marginal effects
```

```
## lm(formula = pot ~ ph + hard + solid + chlor + sulf + cond +     carbon + thm + turb, data = data2)
```

```
##        ph      hard     solid    chlor       sulf       cond    carbon
##  0.006283 -8.73e-06 2.371e-06 0.006897 -9.925e-05 -9.216e-05 -0.002142
##       thm    turb
##  0.0002882 0.01406
```

```
margins(probit)
```

```
## Average marginal effects
```

```
## glm(formula = pot ~ ph + hard + solid + chlor + sulf + cond +     carbon + thm + turb, family = bino
```

```
##        ph       hard     solid    chlor       sulf     cond    carbon       thm
##  0.006004 -1.374e-07 2.359e-06 0.006792 -8.417e-05 -9.2e-05 -0.002147 0.0002834
##      turb
##  0.01402
```

```
margins(logit)
```

```
## Average marginal effects
```

```
## glm(formula = pot ~ ph + hard + solid + chlor + sulf + cond +     carbon + thm + turb, family = bino
```

```
##         ph       hard    solid    chlor       sulf       cond    carbon
##  0.006269 -8.621e-06 2.36e-06 0.006895 -9.857e-05 -9.239e-05 -0.002137
##        thm    turb
##  0.0002876 0.01408
```

We can conclude that the linear probability model is most likely not a good fit for our data, as it can't capture the nonlinear nature of the model, and the regression output gives R2 = 0.0036, i.e. only 0.36% of the variance in potability is explained by the model. It's harder to distinguish between the probit and logit models, as they give similar regression outputs, but the logit model is likely better, as each of the coefficients for the explanatory variables are higher, meaning that the logit model attributes more of the variance in potability to each variable. In addition, the marginal effects of each of the explanatory variables are more similar for the linear probability model and the logit model.

In conclusion, the logit model is better than the probit and LPM models for predicting the potability of water given this data set.