Homework 5

Mari Sanders

2024-11-22

a)

```
state_data <-
  state.x77 %>% as_tibble() %>% janitor::clean_names()
state_data %>% summary()
```

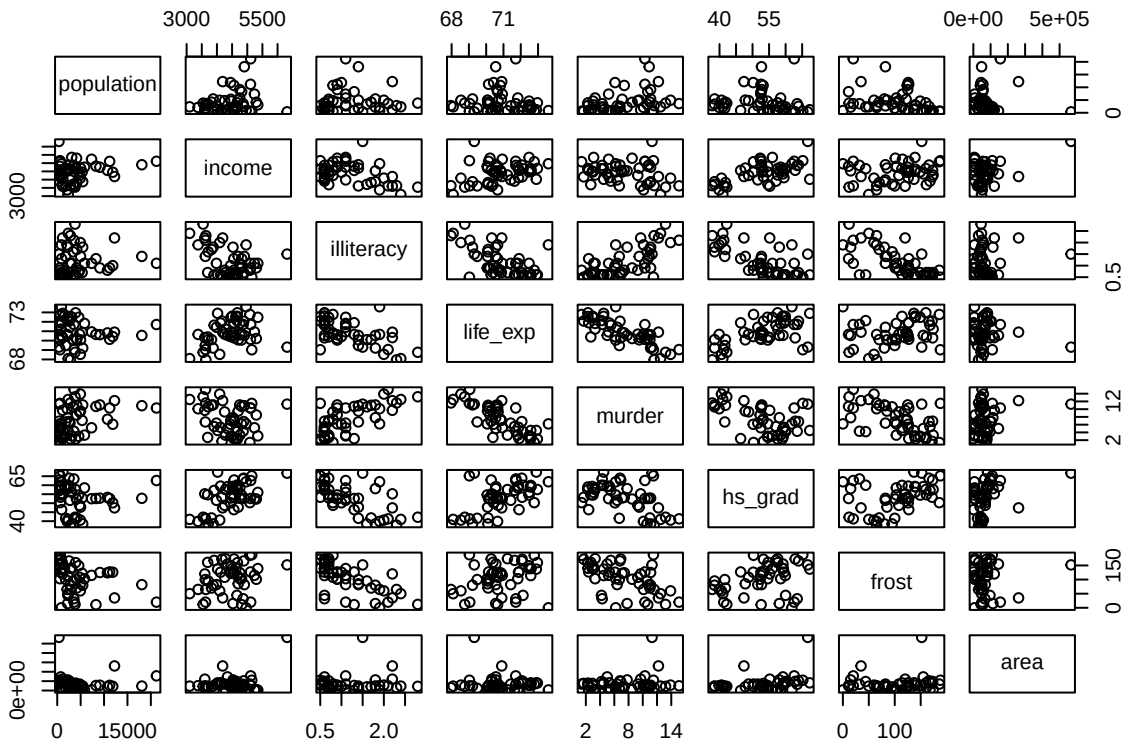
```
##    population        income       illiteracy       life_exp
##  Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
##  1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
##  Median : 2838   Median :4519   Median :0.950   Median :70.67
##  Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88
##  3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
##  Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60
##      murder          hs_grad          frost             area
##  Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
##  1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
##  Median : 6.850   Median :53.25   Median :114.50   Median : 54277
##  Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
##  3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
##  Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

```
state_data %>%
  summarize(population_sd = sd(population, na.rm = TRUE),
            income_sd = sd(income, na.rm = TRUE),
            illiteracy_sd = sd(illiteracy, na.rm = TRUE),
            lifeexpec_sd = sd(life_exp, na.rm = TRUE),
            murder_sd = sd(murder, na.rm = TRUE),
            hsgrad_sd = sd(hs_grad, na.rm = TRUE),
            frost_sd = sd(frost, na.rm = TRUE),
            area_sd = sd(area, na.rm = TRUE))
```

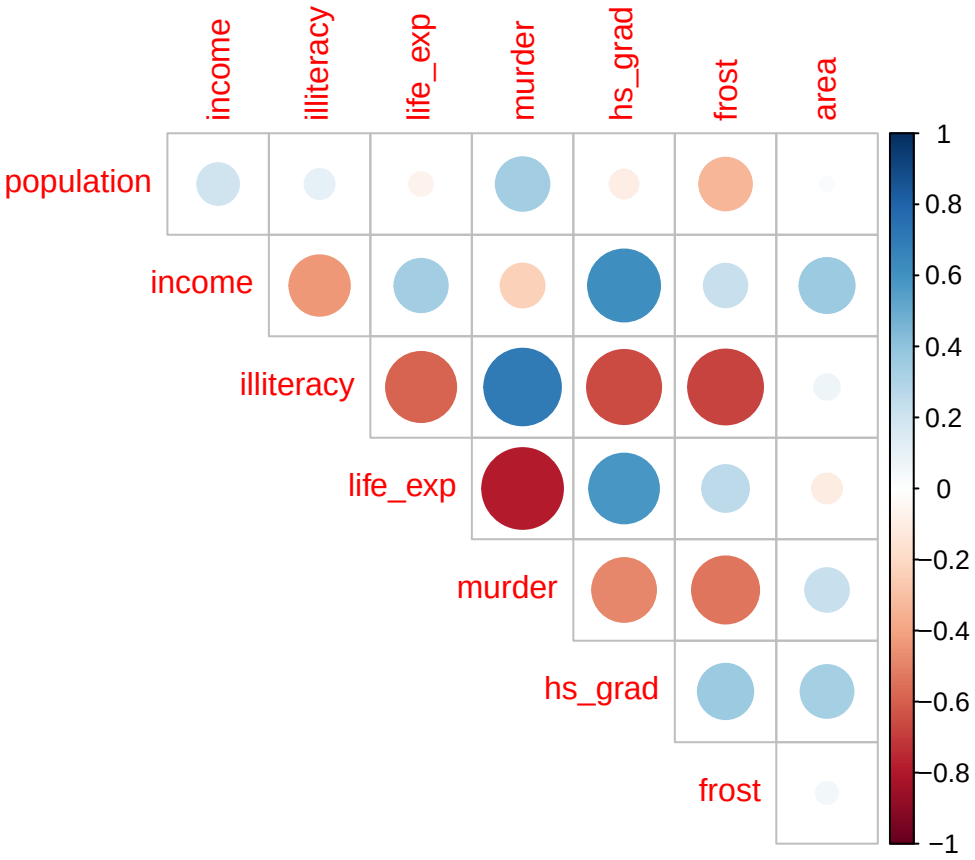
```
## # A tibble: 1 x 8
##   population_sd income_sd illiteracy_sd lifeexpec_sd murder_sd hsgrad_sd
##           <dbl>     <dbl>         <dbl>        <dbl>     <dbl>     <dbl>
## 1         4464.      614.         0.610         1.34      3.69      8.08
## # i 2 more variables: frost_sd <dbl>, area_sd <dbl>
```

b)

```
pairs(state_data)
```



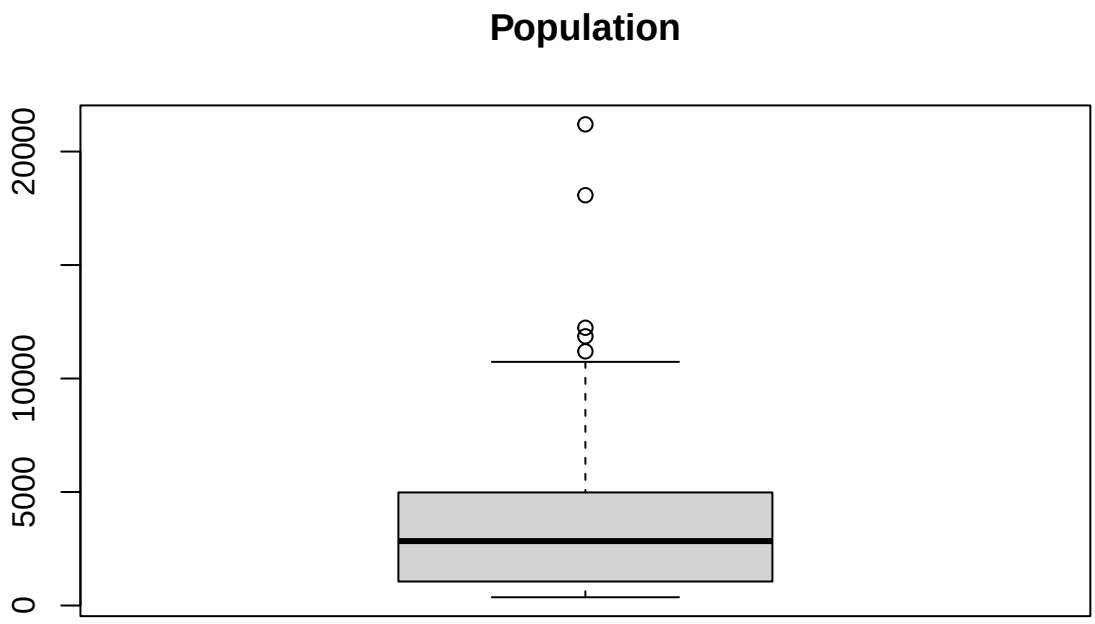

```
corrplot(cor(state_data), type = "upper", diag = FALSE)
```



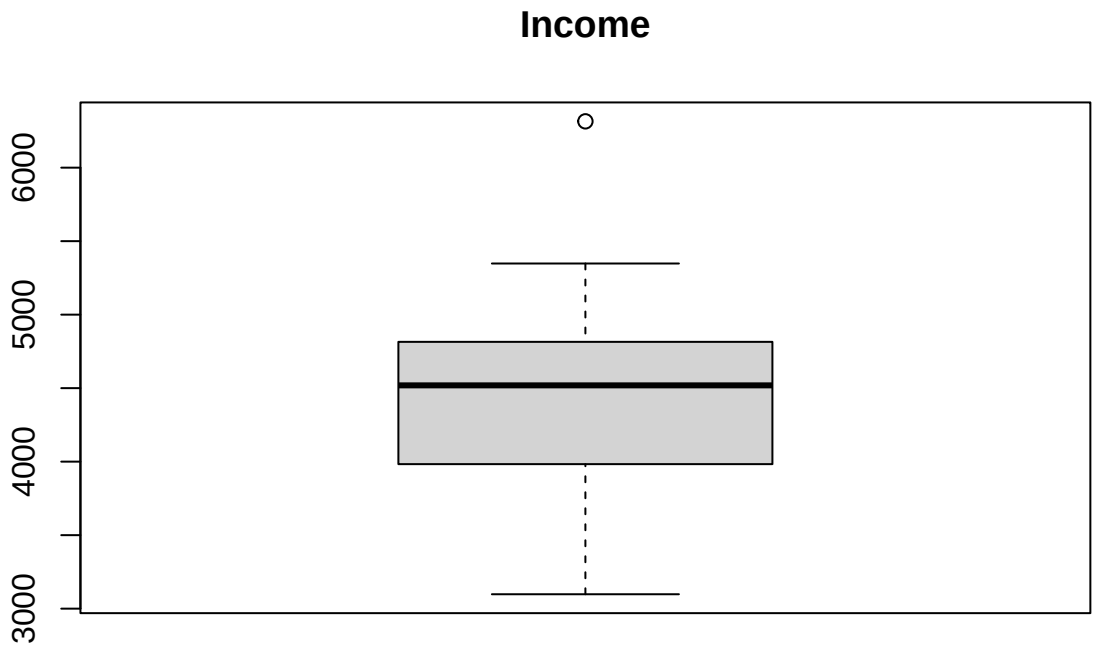

```
par(nfrow = c(2,3))
```

```
## Warning in par(nfrow = c(2, 3)): "nfrow" is not a graphical parameter
```

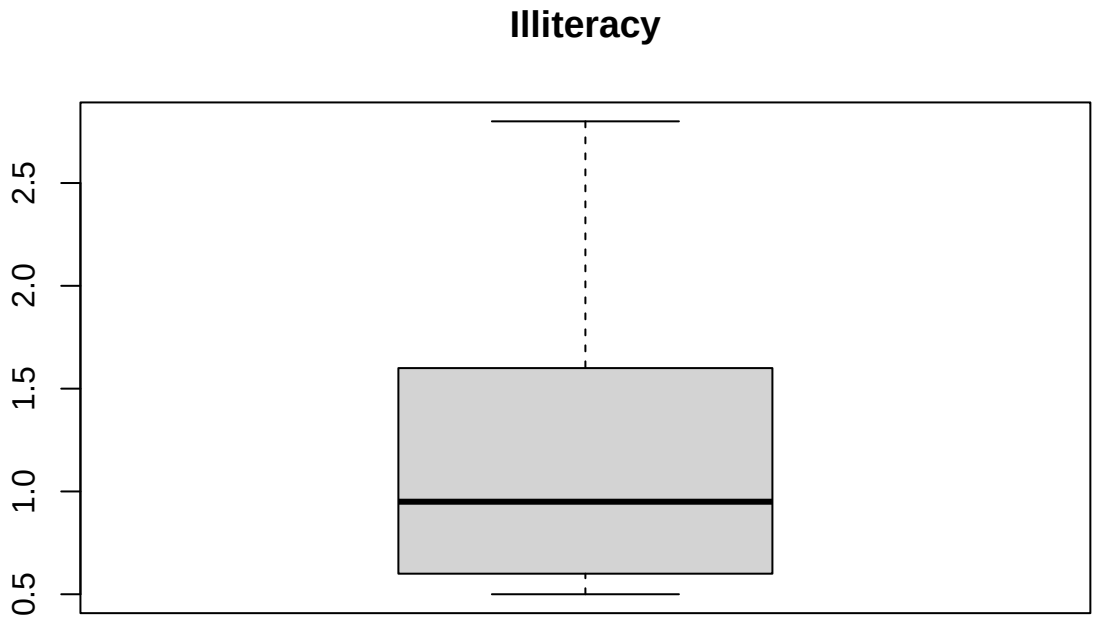
```
boxplot(state_data$population, main = "Population")
```



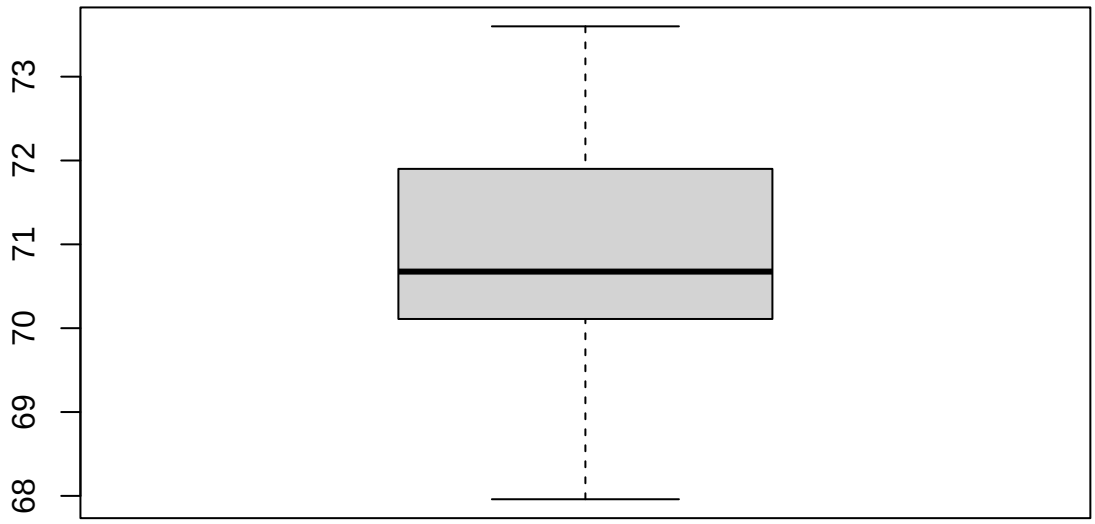

```
boxplot(state_data$income, main = "Income")
```



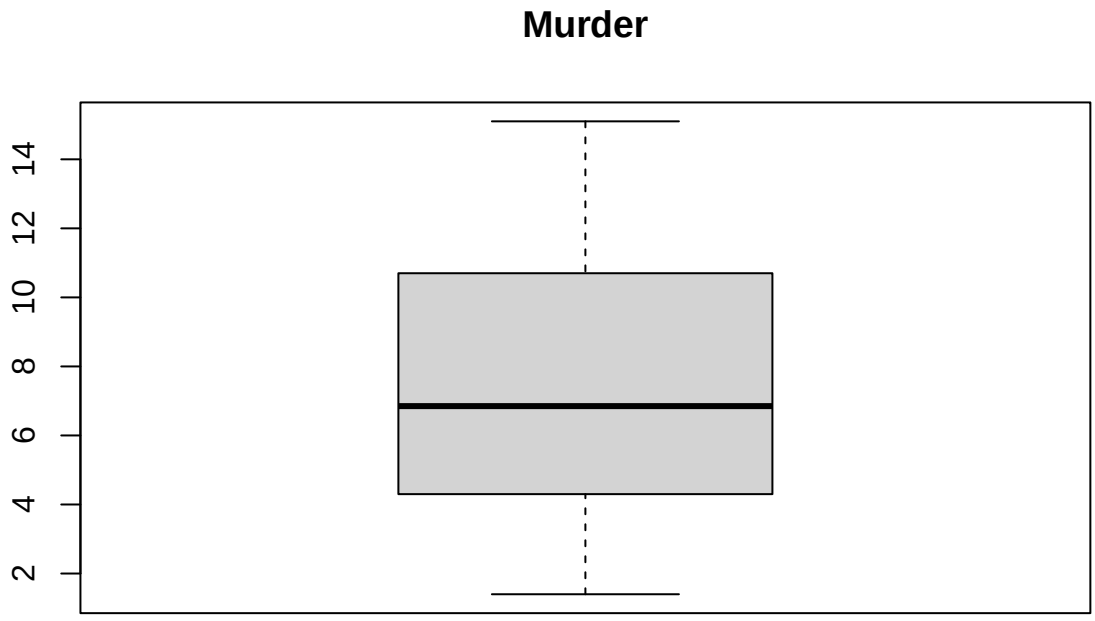

```
boxplot(state_data$illiteracy, main = "Illiteracy")
```



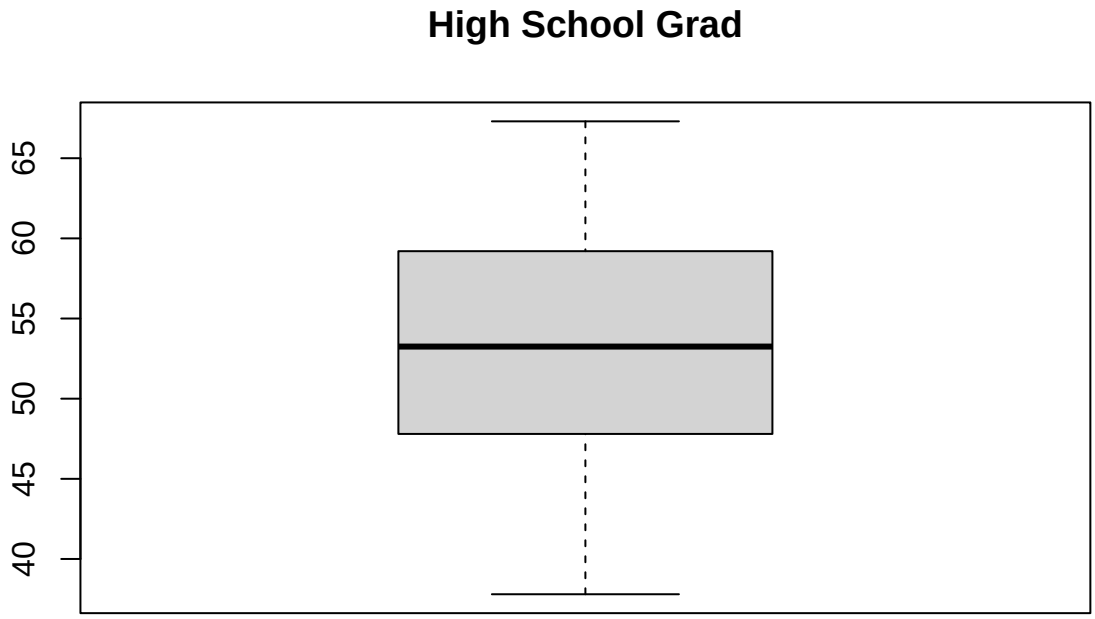

```
boxplot(state_data$life_exp, main = "Life Expectancy")
```



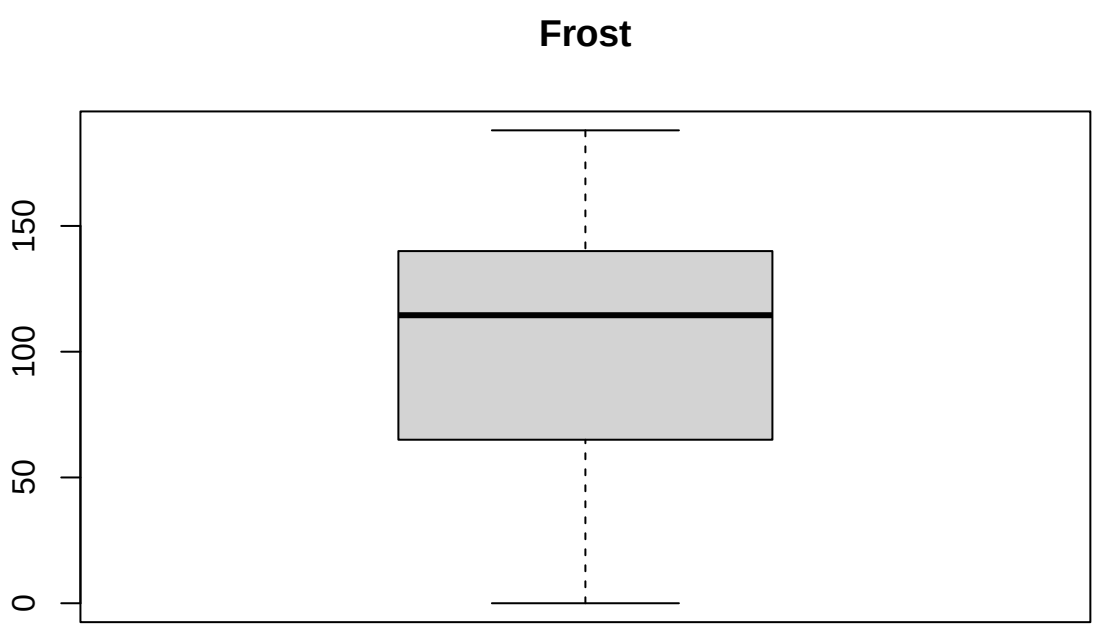

```
boxplot(state_data$murder, main = "Murder")
```



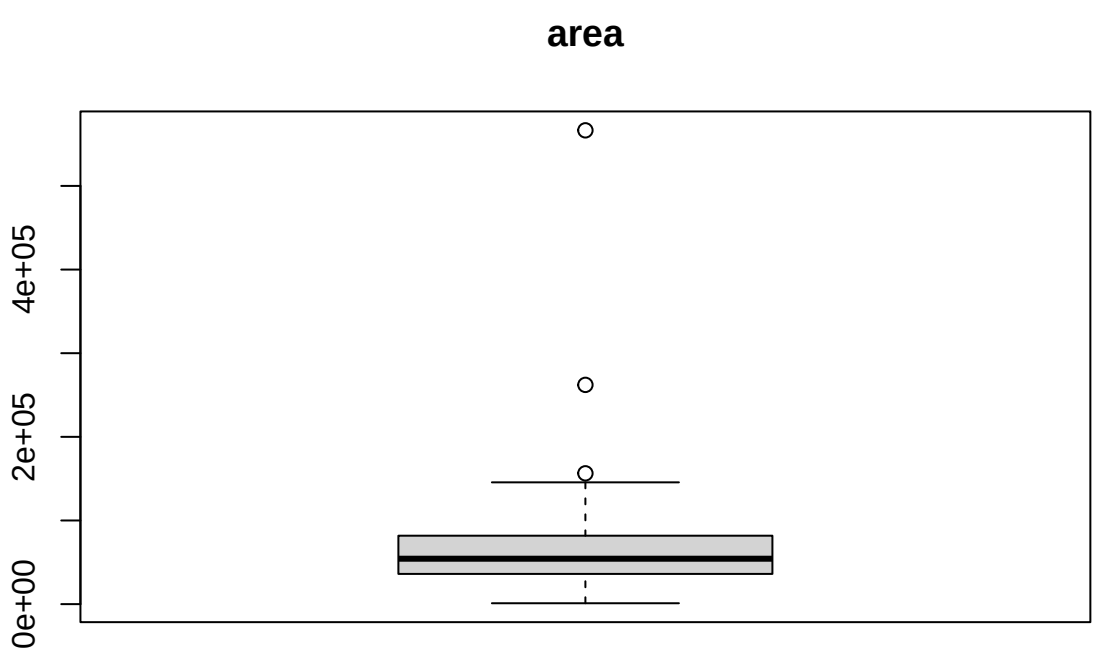

```
boxplot(state_data$hs_grad, main = "High School Grad")
```




```
boxplot(state_data$frost, main = "Frost")
```




```
boxplot(state_data$area, main = "area")
```



Life expectancy, and murder seem to have a relationship, as well as life expectancy and high school grad and life expectancy and illiteracy. There seems to be a slight relationship between life expectancy and frost. It seems like Illiteracy, population, and area are skewed.

c)

```
full_model <- lm(life_exp ~ ., data = state_data)
summary(full_model)
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = state_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## population   5.180e-05  2.919e-05   1.775   0.0832 .
## income      -2.180e-05  2.444e-04  -0.089   0.9293
## illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## hs_grad      4.893e-02  2.332e-02   2.098   0.0420 *
```

```
## frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

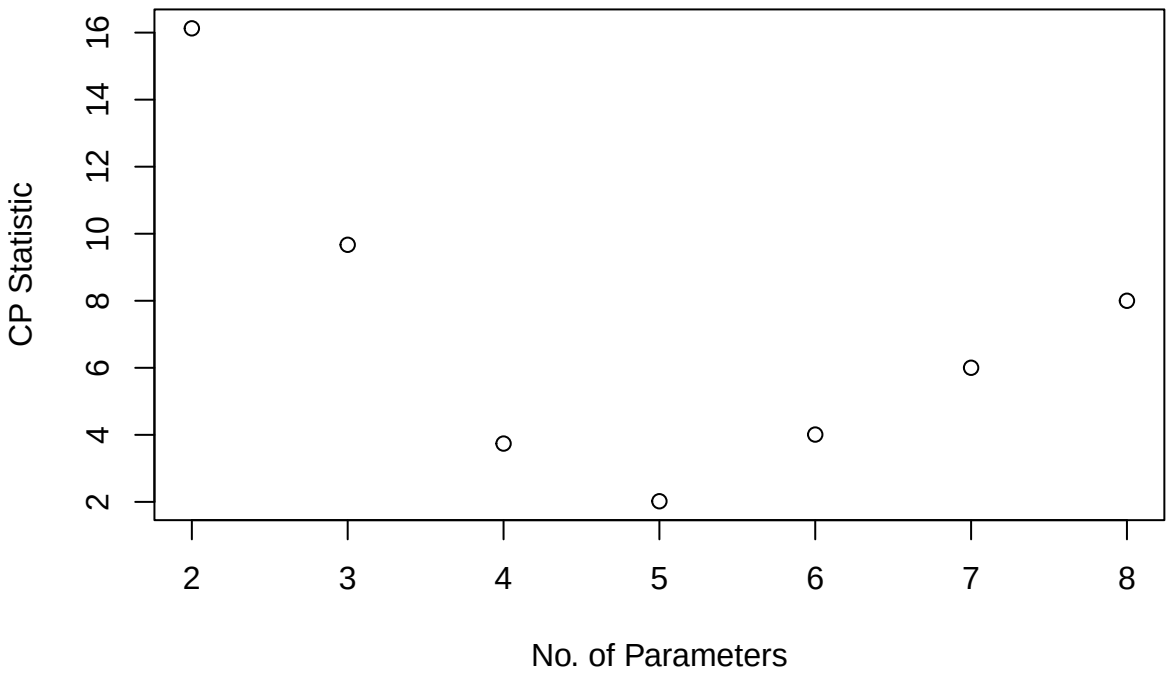
```
subsets = regsubsets(life_exp ~., data = state_data)
subset_res = summary(subsets)
```

d)

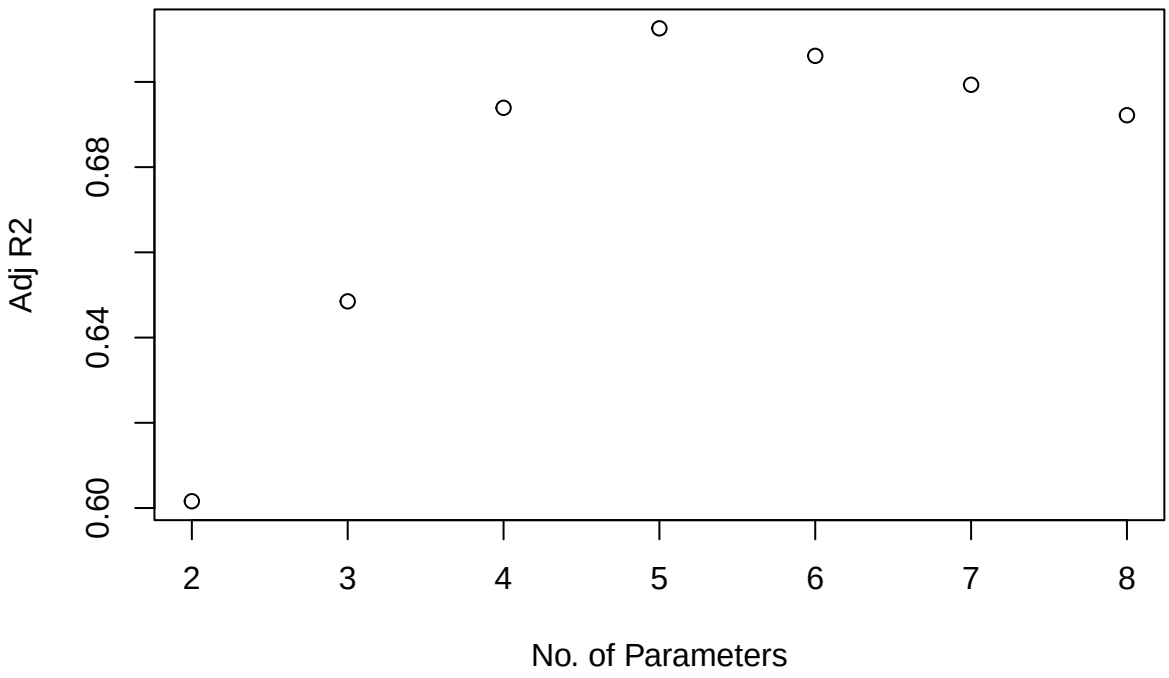
```
subset_res
```

```
## Subset selection object
## Call: regsubsets.formula(life_exp ~ ., data = state_data)
## 7 Variables  (and intercept)
##            Forced in Forced out
## population     FALSE      FALSE
## income         FALSE      FALSE
## illiteracy     FALSE      FALSE
## murder         FALSE      FALSE
## hs_grad        FALSE      FALSE
## frost          FALSE      FALSE
## area           FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          population income illiteracy murder hs_grad frost area
## 1  ( 1 ) " "        " "    " "        "*"    " "     " "   " "
## 2  ( 1 ) " "        " "    " "        "*"    "*"     " "   " "
## 3  ( 1 ) " "        " "    " "        "*"    "*"     "*"   " "
## 4  ( 1 ) "*"        " "    " "        "*"    "*"     "*"   " "
## 5  ( 1 ) "*"        "*"    " "        "*"    "*"     "*"   " "
## 6  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   " "
## 7  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   "*"
```

```
plot(2:8, subset_res$cp, xlab = "No. of Parameters", ylab = "CP Statistic")
```




```
plot(2:8, subset_res$adjr2, xlab = "No. of Parameters", ylab = "Adj R2")
```



Using subsetting, the best model seems to be one that contains population, income, murder, hs_grad, and frost. `life_exp` = `population`β_1 + `income` β_2 + `murder` β_3 + `hs_grad` β_4 + `frost` β_5

e)