Homework 6

Alyssa Vanderbeek (amv2187)

3 December 2018

```
states = state.x77 %>% # load data from faraway package
  as.data.frame() %>%
  janitor::clean_names()
```

1. Explore the dataset and generate appropriate descriptive statistics and relevant graphs for all variables of interest (continuous and categorical) – no test required. Be selective! Even if you create 20 plots, you don't want to show them all.

```
# table of summary stats
states %>%
  skimr::skim_to_list() %>%
  as.data.frame %>%
  dplyr::select(1, 2, 5:11) %>%
  `colnames<-`(c(' ', 'NA', 'Mean', 'Std. Dev.', 'Min', '1st Q', 'Median', '3rd Q', 'Max')) %>%
  knitr::kable()
```

	NA	Mean	Std. Dev.	Min	1st Q	Median	3rd Q	Max
area	0	70735.88	85327.3	1049	36985.25	54277	81162.5	566432
frost	0	104.46	51.98	0	66.25	114.5	139.75	188
hs_grad	0	53.11	8.08	37.8	48.05	53.25	59.15	67.3
illiteracy	0	1.17	0.61	0.5	0.62	0.95	1.58	2.8
income	0	4435.8	614.47	3098	3992.75	4519	4813.5	6315
life_exp	0	70.88	1.34	67.96	70.12	70.67	71.89	73.6
murder	0	7.38	3.69	1.4	4.35	6.85	10.67	15.1
population	0	4246.42	4464.49	365	1079.5	2838.5	4968.5	21198

```
# scatterplot to assess correlation between vars
states %>%
  pairs
```
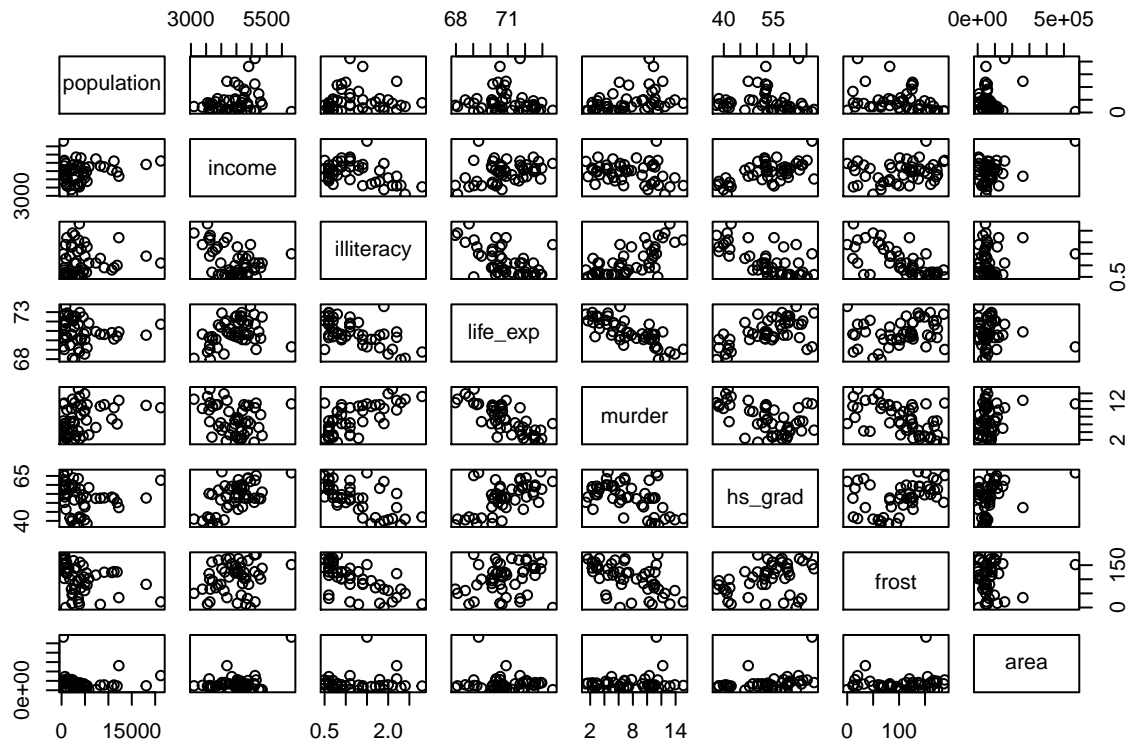
```
# correlation matrix to evaluate what is seen in scatterplots
states %>%
  cor
```

```
##             population      income   illiteracy      life_exp      murder
## population  1.00000000  0.2082276   0.10762237 -0.06805195   0.3436428
## income      0.20822756  1.0000000  -0.43707519  0.34025534  -0.2300776
## illiteracy  0.10762237 -0.4370752   1.00000000 -0.58847793   0.7029752
## life_exp   -0.06805195  0.3402553  -0.58847793  1.00000000  -0.7808458
## murder      0.34364275 -0.2300776   0.70297520 -0.78084575   1.0000000
## hs_grad    -0.09848975  0.6199323  -0.65718861  0.58221620  -0.4879710
## frost      -0.33215245  0.2262822  -0.67194697  0.26206801  -0.5388834
## area        0.02254384  0.3633154   0.07726113 -0.10733194   0.2283902
##                hs_grad       frost         area
## population -0.09848975 -0.3321525   0.02254384
## income      0.61993232  0.2262822   0.36331544
## illiteracy -0.65718861 -0.6719470   0.07726113
## life_exp    0.58221620  0.2620680  -0.10733194
## murder     -0.48797102 -0.5388834   0.22839021
## hs_grad     1.00000000  0.3667797   0.33354187
## frost       0.36677970  1.0000000   0.05922910
## area        0.33354187  0.0592291   1.00000000
```

It looks like murder is correlated both with life expectancy and illiteracy, suggesting that it is a potential confounder. Specifically, murder is positively associated with illiteracy (higher murder rate = higher illiteracy rate) and negatively associated with life expectancy (higher murder rate = lower life expectancy).

After examining the distribution of each vriable in the dataset, I chose to perform a log transformation on the estimates for area size, illiteracy rate, and population size, which were all skewed.
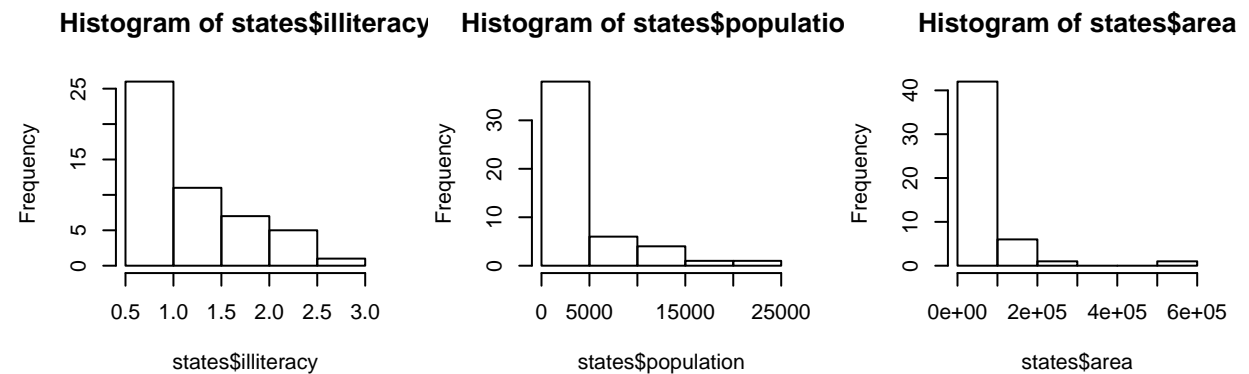
```
states_analysis = states %>%
  mutate(log_area = log(area),
```

```
        log_illiteracy = log(illiteracy),
        log_popn = log(population)) %>%
  dplyr::select(-area, -population, -illiteracy)

par(mfrow = c(2, 3))
hist(states$illiteracy)
hist(states$population)
hist(states$area)
hist(states_analysis$log_illiteracy)
hist(states_analysis$log_popn)
hist(states_analysis$log_area)
```

Histogram of states$illiteracy   **Histogram of states$populatio**   **Histogram of states$area**

togram of states_analysis$log_ili istogram of states_analysis$log_l istogram of states_analysis$log_
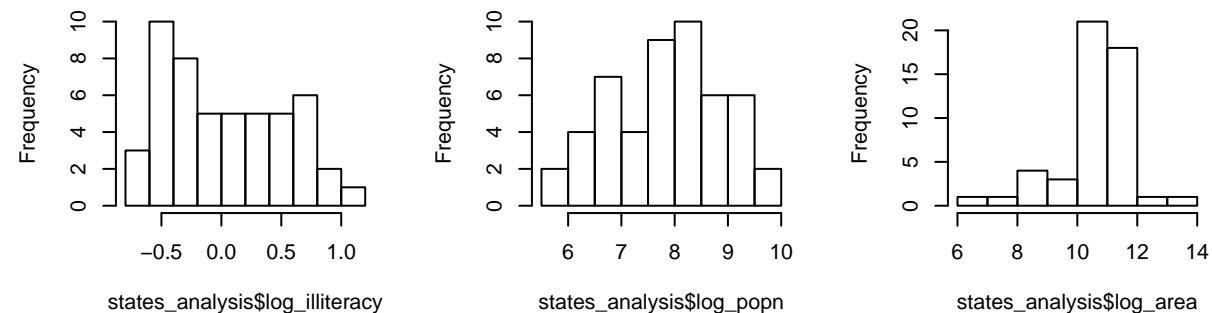
2. Use automatic procedures to find a 'best subset' of the full model. Present the results and comment on the following

```
## backwards elimination
summary(lm(life_exp ~ ., data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44702 -0.42901  0.04546  0.50742  1.68911
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.799e+01  1.798e+00  37.809  < 2e-16 ***
```

3

```
## income        -4.417e-06  2.475e-04  -0.018   0.9858
## murder        -3.114e-01  4.659e-02  -6.684 4.12e-08 ***
## hs_grad         5.482e-02  2.552e-02   2.148   0.0375 *
## frost          -4.669e-03  3.173e-03  -1.471   0.1487
## log_area        7.314e-02  1.102e-01   0.663   0.5107
## log_illiteracy  1.883e-01  4.204e-01   0.448   0.6565
## log_popn        2.537e-01  1.311e-01   1.936   0.0597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7335 on 42 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7014
## F-statistic: 17.45 on 7 and 42 DF,  p-value: 1.368e-10
```

```
summary(lm(life_exp ~ murder + hs_grad + frost + log_area + log_illiteracy + log_popn, data = states_ana
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + log_area +
##     log_illiteracy + log_popn, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.44406 -0.42783  0.04462  0.50722  1.68851
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    67.991653   1.777131  38.259  < 2e-16 ***
## murder         -0.311495   0.045635  -6.826  2.3e-08 ***
## hs_grad         0.054521   0.018818   2.897   0.0059 **
## frost          -0.004684   0.003022  -1.550   0.1284
## log_area        0.073696   0.104455   0.706   0.4843
## log_illiteracy  0.187064   0.409816   0.456   0.6504
## log_popn        0.252730   0.118609   2.131   0.0389 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7249 on 43 degrees of freedom
## Multiple R-squared:  0.7441, Adjusted R-squared:  0.7084
## F-statistic: 20.84 on 6 and 43 DF,  p-value: 2.834e-11
```

```
summary(lm(life_exp ~ murder + hs_grad + frost + log_illiteracy + log_popn, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + log_illiteracy +
##     log_popn, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42070 -0.45738  0.05513  0.53826  1.57824
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    68.428995   1.655984  41.322  < 2e-16 ***
```

```
## murder       -0.296227    0.039947  -7.415 2.83e-09 ***
## hs_grad        0.058095    0.018019   3.224  0.00238 **
## frost         -0.004596    0.003002  -1.531  0.13290
## log_illiteracy 0.140797    0.402220   0.350  0.72797
## log_popn       0.257589    0.117731   2.188  0.03403 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7208 on 44 degrees of freedom
## Multiple R-squared:  0.7411, Adjusted R-squared:  0.7117
## F-statistic: 25.19 on 5 and 44 DF,  p-value: 6.734e-12
```

```
b.fit = lm(life_exp ~ murder + hs_grad + frost + log_popn, data = states_analysis)
summary(b.fit)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost + log_popn,
##     data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41760 -0.43880  0.02539  0.52066  1.63048
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.720810   1.416828  48.503  < 2e-16 ***
## murder      -0.290016   0.035440  -8.183 1.87e-10 ***
## hs_grad      0.054550   0.014758   3.696 0.000591 ***
## frost       -0.005174   0.002482  -2.085 0.042779 *
## log_popn     0.246836   0.112539   2.193 0.033491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7137 on 45 degrees of freedom
## Multiple R-squared:  0.7404, Adjusted R-squared:  0.7173
## F-statistic: 32.09 on 4 and 45 DF,  p-value: 1.17e-12
```

```
## forwards process
summary(lm(life_exp ~ murder, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81690 -0.48139  0.09591  0.39769  2.38691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.97356    0.26997  270.30  < 2e-16 ***
## murder      -0.28395    0.03279   -8.66 2.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.8473 on 48 degrees of freedom
## Multiple R-squared:  0.6097, Adjusted R-squared:  0.6016
## F-statistic: 74.99 on 1 and 48 DF,  p-value: 2.26e-11
```

```
summary(lm(life_exp ~ hs_grad, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ hs_grad, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.01867 -0.67517 -0.07538  0.64483  2.17311
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.73965    1.04748  62.760  < 2e-16 ***
## hs_grad      0.09676    0.01950   4.961  9.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 48 degrees of freedom
## Multiple R-squared:  0.339,  Adjusted R-squared:  0.3252
## F-statistic: 24.61 on 1 and 48 DF,  p-value: 9.196e-06
```

```
summary(lm(life_exp ~ frost, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ frost, data = states_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6515 -0.7852 -0.1183  0.9382  3.4284
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.171631   0.418883 167.521   <2e-16 ***
## frost        0.006768   0.003597   1.881    0.066 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.309 on 48 degrees of freedom
## Multiple R-squared:  0.06868,    Adjusted R-squared:  0.04928
## F-statistic:  3.54 on 1 and 48 DF,  p-value: 0.06599
```

```
summary(lm(life_exp ~ log_area, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ log_area, data = states_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9618 -0.7841 -0.1655  1.0537  2.4849
```

```
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  72.2098     1.7685  40.831   <2e-16 ***
## log_area     -0.1248     0.1649  -0.757    0.453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.348 on 48 degrees of freedom
## Multiple R-squared:  0.0118, Adjusted R-squared:  -0.008786
## F-statistic: 0.5732 on 1 and 48 DF,  p-value: 0.4527
```
```r
summary(lm(life_exp ~ log_illiteracy, data = states_analysis))
```
```
## 
## Call:
## lm(formula = life_exp ~ log_illiteracy, data = states_analysis)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9536 -0.8010  0.0038  0.6943  3.6527
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     70.9263     0.1579 449.148  < 2e-16 ***
## log_illiteracy  -1.5253     0.3174  -4.806 1.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.114 on 48 degrees of freedom
## Multiple R-squared:  0.3249, Adjusted R-squared:  0.3108
## F-statistic:  23.1 on 1 and 48 DF,  p-value: 1.555e-05
```
```r
summary(lm(life_exp ~ log_popn, data = states_analysis))
```
```
## 
## Call:
## lm(formula = life_exp ~ log_popn, data = states_analysis)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.90739 -0.70580 -0.05555  1.05171  2.56688
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.9860     1.4665  49.086   <2e-16 ***
## log_popn     -0.1408     0.1849  -0.762     0.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.348 on 48 degrees of freedom
## Multiple R-squared:  0.01194,    Adjusted R-squared:  -0.008646
## F-statistic:  0.58 on 1 and 48 DF,  p-value: 0.4501
```
```r
# murder has lowest p-val. Start adding secondary vars
summary(lm(life_exp ~ murder + hs_grad, data = states_analysis))
```

```
## 
## Call:
## lm(formula = life_exp ~ murder + hs_grad, data = states_analysis)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66758 -0.41801  0.05602  0.55913  2.05625
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.29708    1.01567  69.213  < 2e-16 ***
## murder      -0.23709    0.03529  -6.719 2.18e-08 ***
## hs_grad      0.04389    0.01613   2.721  0.00909 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7959 on 47 degrees of freedom
## Multiple R-squared:  0.6628, Adjusted R-squared:  0.6485
## F-statistic:   46.2 on 2 and 47 DF,  p-value: 8.016e-12
```

```r
summary(lm(life_exp ~ murder + frost, data = states_analysis))
```

```
## 
## Call:
## lm(formula = life_exp ~ murder + frost, data = states_analysis)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76265 -0.55772  0.07629  0.44294  1.73190
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.900325   0.500294 147.714  < 2e-16 ***
## murder      -0.327777   0.037505  -8.739 2.05e-11 ***
## frost       -0.005776   0.002664  -2.169   0.0352 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8164 on 47 degrees of freedom
## Multiple R-squared:  0.6452, Adjusted R-squared:  0.6301
## F-statistic: 42.74 on 2 and 47 DF,  p-value: 2.655e-11
```

```r
summary(lm(life_exp ~ murder + log_area, data = states_analysis))
```

```
## 
## Call:
## lm(formula = life_exp ~ murder + log_area, data = states_analysis)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83475 -0.54368  0.06741  0.44642  2.66277
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.43176    1.10270  64.779  < 2e-16 ***
```

```
## murder        -0.29845    0.03395  -8.790 1.73e-11 ***
## log_area       0.15463    0.10730   1.441    0.156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.838 on 47 degrees of freedom
## Multiple R-squared:  0.6262, Adjusted R-squared:  0.6103
## F-statistic: 39.37 on 2 and 47 DF,  p-value: 9.038e-11
```
```r
summary(lm(life_exp ~ murder + log_illiteracy, data = states_analysis))
```
```
##
## Call:
## lm(formula = life_exp ~ murder + log_illiteracy, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80926 -0.47370  0.09391  0.41515  2.48964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    72.87888    0.35332 206.269  < 2e-16 ***
## murder         -0.27051    0.04598  -5.883 4.05e-07 ***
## log_illiteracy -0.14233    0.33839  -0.421    0.676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8547 on 47 degrees of freedom
## Multiple R-squared:  0.6112, Adjusted R-squared:  0.5946
## F-statistic: 36.94 on 2 and 47 DF,  p-value: 2.286e-10
```
```r
summary(lm(life_exp ~ murder + log_popn, data = states_analysis))
```
```
##
## Call:
## lm(formula = life_exp ~ murder + log_popn, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70906 -0.42411  0.09279  0.45195  2.63524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.16651    0.89471  79.541  < 2e-16 ***
## murder      -0.30972    0.03394  -9.125 5.64e-12 ***
## log_popn     0.25399    0.12029   2.111   0.0401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8184 on 47 degrees of freedom
## Multiple R-squared:  0.6435, Adjusted R-squared:  0.6284
## F-statistic: 42.42 on 2 and 47 DF,  p-value: 2.968e-11
```
```r
# murder + hs_grad
summary(lm(life_exp ~ murder + hs_grad + frost, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost, data = states_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## hs_grad      0.049949   0.015201   3.286  0.00195 **
## frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```
```r
summary(lm(life_exp ~ murder + hs_grad + log_area, data = states_analysis))
```
```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_area, data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68415 -0.36849  0.06854  0.54491  2.16838
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.00002    1.22868  56.972  < 2e-16 ***
## murder      -0.24513    0.04005  -6.121 1.91e-07 ***
## hs_grad      0.04069    0.01783   2.282   0.0272 *
## log_area     0.04935    0.11268   0.438   0.6634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8028 on 46 degrees of freedom
## Multiple R-squared:  0.6642, Adjusted R-squared:  0.6423
## F-statistic: 30.34 on 3 and 46 DF,  p-value: 5.681e-11
```
```r
summary(lm(life_exp ~ murder + hs_grad + log_illiteracy, data = states_analysis))
```
```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_illiteracy, data = states_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6515 -0.5034  0.1132  0.5889  1.6629
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69.88243    1.07409  65.062  < 2e-16 ***
## murder         -0.26505    0.04271  -6.206 1.42e-07 ***
## hs_grad         0.05533    0.01889   2.930  0.00527 **
## log_illiteracy  0.42558    0.36902   1.153  0.25476
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7931 on 46 degrees of freedom
## Multiple R-squared:  0.6723, Adjusted R-squared:  0.6509
## F-statistic: 31.46 on 3 and 46 DF,  p-value: 3.264e-11
```

```r
summary(lm(life_exp ~ murder + hs_grad + log_popn, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn, data = states_analysis)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.50380 -0.59250  0.06207  0.42032  2.31066
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.52120    1.34101  50.351  < 2e-16 ***
## murder      -0.26121    0.03380  -7.727 7.48e-10 ***
## hs_grad      0.05187    0.01523   3.406  0.00138 **
## log_popn     0.32172    0.11046   2.912  0.00552 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7392 on 46 degrees of freedom
## Multiple R-squared:  0.7153, Adjusted R-squared:  0.6968
## F-statistic: 38.53 on 3 and 46 DF,  p-value: 1.32e-12
```

```r
# murder + hs_grad + log_popn
summary(lm(life_exp ~ murder + hs_grad + log_popn + frost, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn + frost,
##     data = states_analysis)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.41760 -0.43880  0.02539  0.52066  1.63048
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.720810   1.416828  48.503  < 2e-16 ***
## murder      -0.290016   0.035440  -8.183 1.87e-10 ***
## hs_grad      0.054550   0.014758   3.696 0.000591 ***
## log_popn     0.246836   0.112539   2.193 0.033491 *
## frost       -0.005174   0.002482  -2.085 0.042779 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7137 on 45 degrees of freedom
## Multiple R-squared:  0.7404, Adjusted R-squared:  0.7173
## F-statistic: 32.09 on 4 and 45 DF,  p-value: 1.17e-12
```

```r
summary(lm(life_exp ~ murder + hs_grad + log_popn + log_area, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn + log_area,
##     data = states_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5157 -0.5875  0.0417  0.4582  2.3833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.34149    1.47257  45.730  < 2e-16 ***
## murder      -0.26638    0.03798  -7.014  9.7e-09 ***
## hs_grad      0.04971    0.01688   2.945  0.00509 **
## log_popn     0.31979    0.11174   2.862  0.00637 **
## log_area     0.03263    0.10495   0.311  0.75732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7466 on 45 degrees of freedom
## Multiple R-squared:  0.7159, Adjusted R-squared:  0.6907
## F-statistic: 28.36 on 4 and 45 DF,  p-value: 8.603e-12
```

```r
summary(lm(life_exp ~ murder + hs_grad + log_popn + log_illiteracy, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn + log_illiteracy,
##     data = states_analysis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48154 -0.57675  0.00752  0.51252  1.87366
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    66.98373    1.38078  48.511  < 2e-16 ***
## murder         -0.29334    0.04049  -7.244 4.43e-09 ***
## hs_grad         0.06497    0.01771   3.669 0.000643 ***
## log_popn        0.32985    0.10946   3.013 0.004231 **
## log_illiteracy  0.47971    0.34081   1.408 0.166130
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7314 on 45 degrees of freedom
## Multiple R-squared:  0.7273, Adjusted R-squared:  0.7031
## F-statistic: 30.01 on 4 and 45 DF,  p-value: 3.476e-12
```

```
# murder + hs_grad + log_popn + frost
summary(lm(life_exp ~ murder + hs_grad + log_popn + frost + log_area, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn + frost +
##     log_area, data = states_analysis)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.43763 -0.46147  0.00721  0.48073  1.74473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.414712   1.502603  45.531  < 2e-16 ***
## murder      -0.301875   0.040110  -7.526 1.95e-09 ***
## hs_grad      0.050302   0.016243   3.097   0.0034 **
## log_popn     0.239312   0.113870   2.102   0.0413 *
## frost       -0.005424   0.002528  -2.146   0.0374 *
## log_area     0.066067   0.102178   0.647   0.5213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7184 on 44 degrees of freedom
## Multiple R-squared:  0.7429, Adjusted R-squared:  0.7136
## F-statistic: 25.42 on 5 and 44 DF,  p-value: 5.833e-12
```

```
summary(lm(life_exp ~ murder + hs_grad + log_popn + frost + log_illiteracy, data = states_analysis))
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn + frost +
##     log_illiteracy, data = states_analysis)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.42070 -0.45738  0.05513  0.53826  1.57824
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    68.428995   1.655984  41.322  < 2e-16 ***
## murder         -0.296227   0.039947  -7.415 2.83e-09 ***
## hs_grad         0.058095   0.018019   3.224  0.00238 **
## log_popn        0.257589   0.117731   2.188  0.03403 *
## frost          -0.004596   0.003002  -1.531  0.13290
## log_illiteracy  0.140797   0.402220   0.350  0.72797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7208 on 44 degrees of freedom
## Multiple R-squared:  0.7411, Adjusted R-squared:  0.7117
## F-statistic: 25.19 on 5 and 44 DF,  p-value: 6.734e-12
```

```
f.fit = lm(life_exp ~ murder + hs_grad + log_popn + frost, data = states_analysis)
summary(f.fit)
```

```
## 
## Call:
## lm(formula = life_exp ~ murder + hs_grad + log_popn + frost,
##     data = states_analysis)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41760 -0.43880  0.02539  0.52066  1.63048
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.720810   1.416828  48.503  < 2e-16 ***
## murder      -0.290016   0.035440  -8.183 1.87e-10 ***
## hs_grad      0.054550   0.014758   3.696 0.000591 ***
## log_popn     0.246836   0.112539   2.193 0.033491 *
## frost       -0.005174   0.002482  -2.085 0.042779 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7137 on 45 degrees of freedom
## Multiple R-squared:  0.7404, Adjusted R-squared:  0.7173
## F-statistic: 32.09 on 4 and 45 DF,  p-value: 1.17e-12
## Stepwise
step.fit = step(lm(life_exp ~ ., data = states_analysis))

## Start:  AIC=-23.71
## life_exp ~ income + murder + hs_grad + frost + log_area + log_illiteracy +
##     log_popn
## 
##                  Df Sum of Sq    RSS     AIC
## - income          1    0.0002 22.596 -25.712
## - log_illiteracy  1    0.1079 22.704 -25.475
## - log_area        1    0.2368 22.833 -25.192
## <none>                        22.596 -23.713
## - frost           1    1.1645 23.760 -23.200
## - log_popn        1    2.0155 24.611 -21.441
## - hs_grad         1    2.4822 25.078 -20.502
## - murder          1   24.0347 46.631  10.512
## 
## Step:  AIC=-25.71
## life_exp ~ murder + hs_grad + frost + log_area + log_illiteracy +
##     log_popn
## 
##                  Df Sum of Sq    RSS      AIC
## - log_illiteracy  1    0.1095 22.705 -27.4708
## - log_area        1    0.2616 22.858 -27.1370
## <none>                        22.596 -25.7125
## - frost           1    1.2628 23.859 -24.9936
## - log_popn        1    2.3859 24.982 -22.6937
## - hs_grad         1    4.4112 27.007 -18.7959
## - murder          1   24.4834 47.079   8.9907
## 
## Step:  AIC=-27.47
## life_exp ~ murder + hs_grad + frost + log_area + log_popn
```

```
## 
##              Df Sum of Sq    RSS     AIC
## - log_area  1     0.2157 22.921 -28.998
## <none>                     22.705 -27.471
## - log_popn  1     2.2792 24.985 -24.688
## - frost     1     2.3760 25.082 -24.495
## - hs_grad   1     4.9491 27.655 -19.612
## - murder    1    29.2296 51.935  11.899
## 
## Step:  AIC=-29
## life_exp ~ murder + hs_grad + frost + log_popn
## 
##              Df Sum of Sq    RSS     AIC
## <none>                     22.921 -28.998
## - frost     1      2.214 25.135 -26.387
## - log_popn  1      2.450 25.372 -25.920
## - hs_grad   1      6.959 29.881 -17.741
## - murder    1     34.109 57.031  14.578
```

All automatic processes conclude the same model, using percent increase in population size (log(population)), rate of high school graduation (hs_grad), murder rate per 100,000 (murder), and average number of days annually with temperatures below freezing (frost) as predictors of life expectancy.

'Frost' is the least significant predictor, with a p-value of 0.043. No variables were seen to be a "close call" at the 5% significance level.

There is a correlation between illiteracy (with and without log transformation) and high school graduation rate (-0.6571886), but my model includes only high school graduation rate as a predictor.

3. Use criterion-based procedures studied in class to guide your selection of the 'best subset'. Summarize your results (tabular or graphical)

```r
leaps::leaps(x = states_analysis[, c(1, 3:8)], y = states_analysis$life_exp, nbest = 2, method = "Cp")
```

```
## $which
##       1     2     3     4     5     6     7
## 1 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 2 FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## 2 FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 3 FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE
## 3 FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
## 4 FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## 4 FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## 5 FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 5 FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 6 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
## [6] "5"           "6"           "7"
## 
## $size
##  [1] 2 2 3 3 4 4 5 5 6 6 7 7 8
## 
```

15

```
## $Cp
##  [1] 18.054999 62.491135 11.335656 14.228518  4.720403  5.159514  2.604836
##  [8]  4.750158  4.203829  4.486517  6.000319  6.200619  8.000000
```

```
leaps::leaps(x = states_analysis[, c(1, 3:8)], y = states_analysis$life_exp, nbest = 2, method = "adjr2
```

```
## $which
##        1     2     3     4     5     6     7
## 1 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## 1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 2 FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## 2 FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 3 FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE
## 3 FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
## 4 FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE
## 4 FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## 5 FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 5 FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 6 FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
## 7  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##
## $label
## [1] "(Intercept)" "1"           "2"           "3"           "4"
## [6] "5"           "6"           "7"
##
## $size
##  [1] 2 2 3 3 4 4 5 5 6 6 7 7 8
##
## $adjr2
##  [1] 0.6015893 0.3252044 0.6484991 0.6301232 0.6967729 0.6939230 0.7173392
##  [8] 0.7031061 0.7136360 0.7117179 0.7083894 0.7069987 0.7014485
```