Homework 6

Alyssa Vanderbeek (amv2187)

3 December 2018

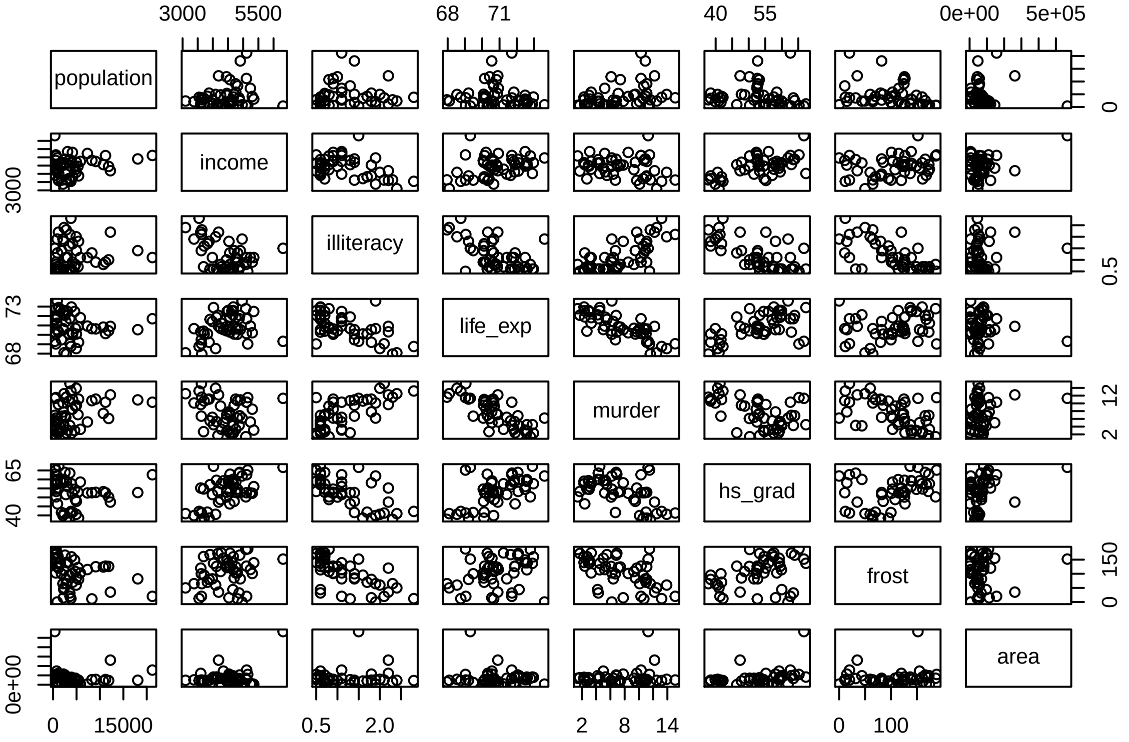
```
states = state.x77 %>% # load data from faraway package
  as.data.frame() %>%
  janitor::clean_names()
```

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
```



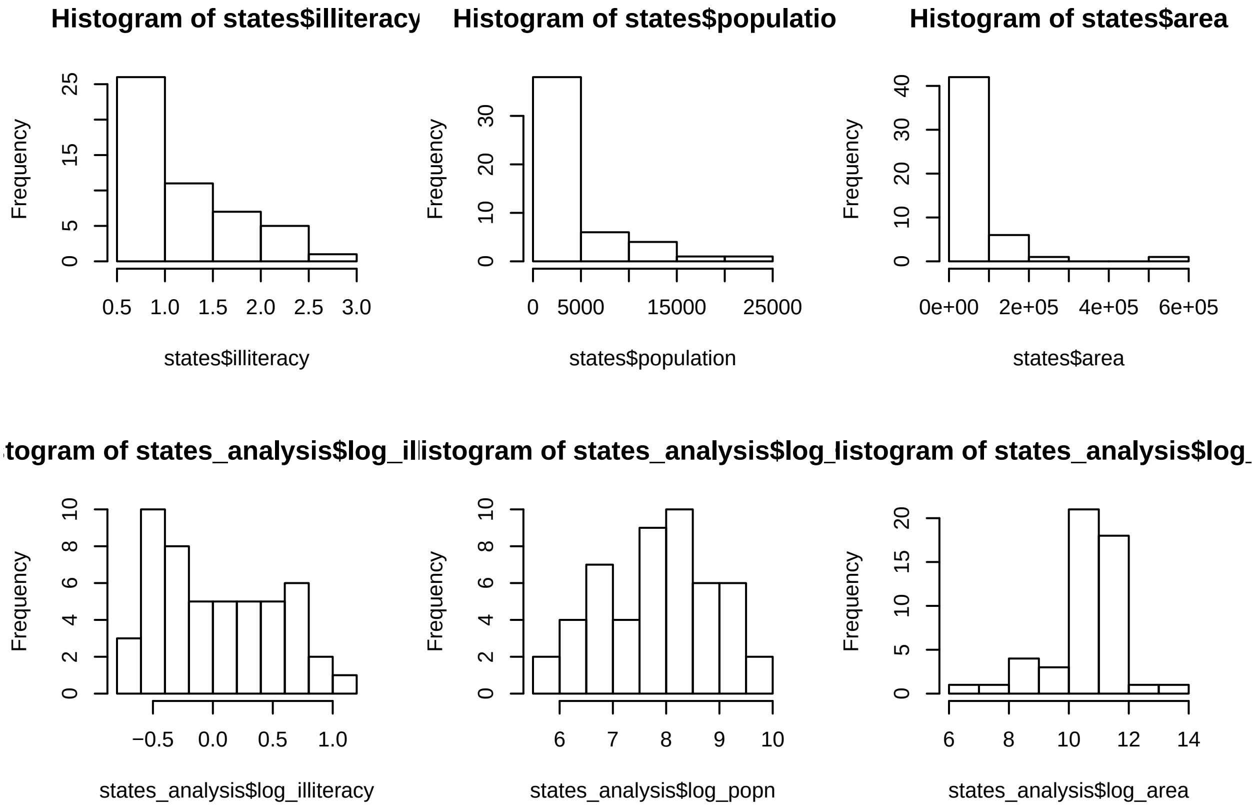
```
# correlation matrix to evaluate what is seen in scatterplots
states %>%
  cor
```

```
##              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
```

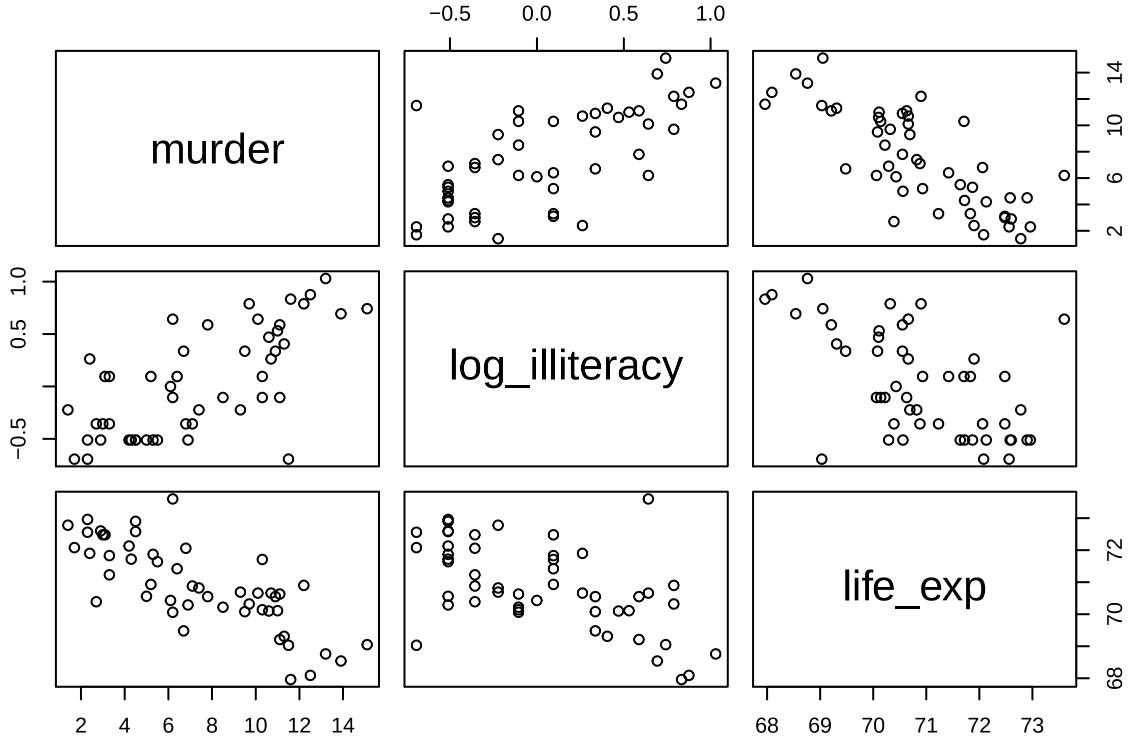
```
# It looks like murder is correlated both with life expectancy and illiteracy, suggesting that it is a
```

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

```r
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)
```



```r
# check correlation between murder, illiteracy, and life expectancy after transforming illiteracy rate
states_analysis %>%
  dplyr::select(murder, log_illiteracy, life_exp) %>%
  pairs
```



```
states_analysis %>%
  dplyr::select(murder, log_illiteracy, life_exp) %>%
  cor
```

```
##                    murder log_illiteracy   life_exp
## murder          1.0000000      0.6947320 -0.7808458
## log_illiteracy  0.6947320      1.0000000 -0.5699943
## life_exp       -0.7808458     -0.5699943  1.0000000
```

```
# Transforming the illiteracy rate reduces the correlation with murder rate slightly (from 0.7 to 0.69)
```

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 ***
## 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
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
```