biostatistical methods homework 5

```
library(knitr)
library(tidyverse)
library(faraway)
library(broom)
```

R dataset 'state.x77' from library(faraway) contains information on 50 states from 1970s collected by US Census Bureau. The goal is to predict 'life expectancy' using a combination of remaining variables.

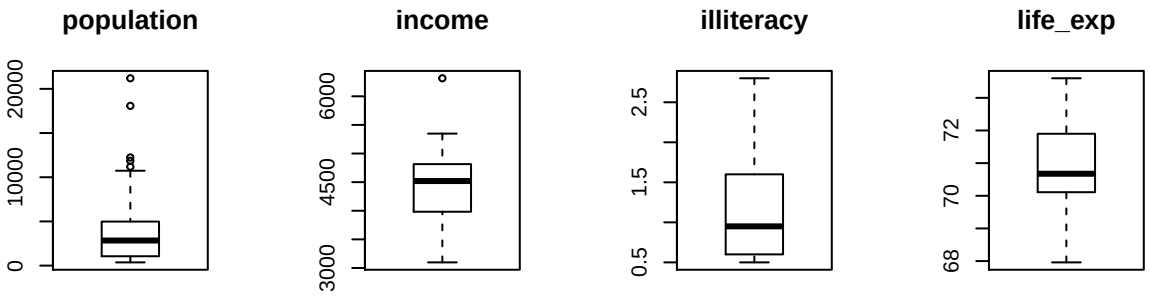
```
life_data = as.data.frame(state.x77) %>%
  janitor::clean_names()
```

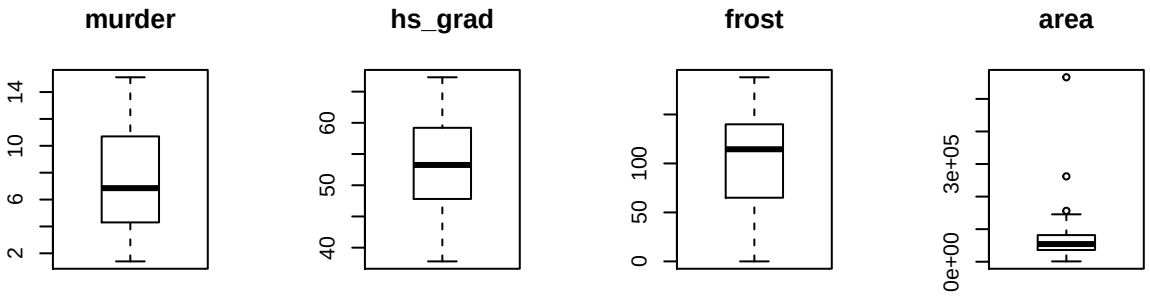
1. Explore the dataset and generate appropriate descriptive statistics and relevant graphs

```
mean_and_sd = function(x) {

  if (!is.numeric(x)) {
    stop("Argument x should be numeric")
  } else if (length(x) == 1) {
    stop("Cannot be computed for length 1 vectors")
  }

  mean_x = mean(x)
  sd_x = sd(x)
  tibble(
    mean = mean_x,
    sd = sd_x
  )
}
```

```
attach(life_data)
```

```
par(mfrow = c(2, 4))
boxplot(population, main = 'population')
boxplot(income,main = 'income' )
boxplot(illiteracy, main = 'illiteracy')
boxplot(life_exp, main = 'life_exp')
boxplot(murder, main = 'murder')
boxplot(hs_grad, main = 'hs_grad')
boxplot(frost, main = 'frost')
boxplot(area, main = 'area')
```





Population

```
summary(population)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     365    1080    2838    4246    4968   21198
```

Income

```
summary(income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3098    3993    4519    4436    4814    6315
```

Illiteracy

```
summary(illiteracy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   0.625   0.950   1.170   1.575   2.800
```

Life Exp

```
summary(life_exp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   67.96   70.12   70.67   70.88   71.89   73.60
```

Murder

```
summary(murder)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   4.350   6.850   7.378  10.675  15.100
```

HS Grad

```
summary(hs_grad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   37.80   48.05   53.25   53.11   59.15   67.30
```

Frost

```
summary(frost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   66.25  114.50  104.46  139.75  188.00
```

Area

```
summary(area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1049   36985   54277   70736   81162  566432
```

2. Use automatic procedures to find a 'best subset' of the full model. Present the results and comment on the following:

```
life_data.fit <- lm(life_exp ~., data=life_data)
summary(life_data.fit)
```

```
##
## Call:
## lm(formula = life_exp ~ ., data = life_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
```

Backward elimination

```
step1<-update(life_data.fit, . ~ . -area)
summary(step1)
```

```
##
## Call:
## lm(formula = life_exp ~ population + income + illiteracy + murder +
##     hs_grad + frost, data = life_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49047 -0.52533 -0.02546  0.57160  1.50374
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
## population   5.188e-05  2.879e-05   1.802   0.0785 .
## income      -2.444e-05  2.343e-04  -0.104   0.9174
## illiteracy   2.846e-02  3.416e-01   0.083   0.9340
## murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
## hs_grad      4.847e-02  2.067e-02   2.345   0.0237 *
## frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7361 on 43 degrees of freedom
## Multiple R-squared:  0.7361, Adjusted R-squared:  0.6993
## F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11
```

```
step2<-update(step1, . ~ . -illiteracy)
summary(step2)
```

```
##
## Call:
## lm(formula = life_exp ~ population + income + murder + hs_grad +
##     frost, data = life_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4892 -0.5122 -0.0329  0.5645  1.5166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
## population   5.115e-05  2.709e-05   1.888   0.0657 .
## income      -2.477e-05  2.316e-04  -0.107   0.9153
## murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
## hs_grad      4.776e-02  1.859e-02   2.569   0.0137 *
## frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7277 on 44 degrees of freedom
## Multiple R-squared:  0.7361, Adjusted R-squared:  0.7061
## F-statistic: 24.55 on 5 and 44 DF,  p-value: 1.019e-11
```

```
step3<-update(step2, . ~ . -income)
summary(step3)
```

```
##
## Call:
## lm(formula = life_exp ~ population + murder + hs_grad + frost,
##     data = life_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## population   5.014e-05  2.512e-05   1.996  0.05201 .
## murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## hs_grad      4.658e-02  1.483e-02   3.142  0.00297 **
## frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

```
step4<-update(step3, . ~ . -population)
summary(step4)
```

```
##
## Call:
## lm(formula = life_exp ~ murder + hs_grad + frost, data = life_data)
##
## 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
```

Using backward elimination, the final model contains murder, hs_grad and frost.

Forward elimination

```
fit1 <- lm(life_exp ~ population, data=life_data)
tidy(fit1)
```

```
## # A tibble: 2 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.0       0.265       267.    7.90e-78
## 2 population   -0.0000205 0.0000433    -0.473 6.39e- 1
```

```
fit2 <- lm(life_exp ~ income, data=life_data)
tidy(fit2)
```

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 67.6       1.33         50.9  1.98e-43
## 2 income       0.000743  0.000297      2.51 1.56e- 2
```

```
fit3 <- lm(life_exp ~ illiteracy, data=life_data)
tidy(fit3)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    72.4      0.338    214.   3.47e-73
## 2 illiteracy     -1.30     0.257     -5.04 6.97e- 6
```

```
fit4 <- lm(life_exp ~ murder, data=life_data)
tidy(fit4)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   73.0      0.270     270.   4.72e-78
## 2 murder        -0.284    0.0328     -8.66 2.26e-11
```

```
fit5 <- lm(life_exp ~ hs_grad, data=life_data)
tidy(fit5)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  65.7       1.05       62.8  9.92e-48
## 2 hs_grad       0.0968    0.0195      4.96 9.20e- 6
```

```
fit6 <- lm(life_exp ~ frost, data=life_data)
tidy(fit6)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 70.2       0.419      168.   4.33e-68
## 2 frost        0.00677   0.00360      1.88 6.60e- 2
```

```
fit7 <- lm(life_exp ~ area, data=life_data)
tidy(fit7)
```

```
## # A tibble: 2 x 5
##   term            estimate  std.error statistic  p.value
##   <chr>              <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.0        0.249        285.    3.46e-79
## 2 area         -0.00000169 0.00000226    -0.748 4.58e- 1
```

```
forward1<-lm(life_exp~murder, data=life_data)
tidy(forward1)
```

```
## # A tibble: 2 x 5
```

```
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   73.0      0.270     270.   4.72e-78
## 2 murder        -0.284    0.0328     -8.66 2.26e-11
```

```
fit1 <- update(forward1, . ~ . +population)
tidy(fit1)
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  72.9       0.258        282.   1.55e-77
## 2 murder       -0.312     0.0332        -9.42 2.15e-12
## 3 population    0.0000683 0.0000274      2.49 1.64e- 2
```

```
fit2 <- update(forward1, . ~ . +income)
tidy(fit2)
```

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.2       0.967        73.6  3.32e-50
## 2 murder       -0.270     0.0328       -8.21 1.22e-10
## 3 income        0.000370  0.000197      1.88 6.66e- 2
```

```
fit3 <- update(forward1, . ~ . +illiteracy)
tidy(fit3)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   73.0      0.286    256.    1.56e-75
## 2 murder        -0.264    0.0464    -5.69  7.96e- 7
## 3 illiteracy    -0.172    0.281     -0.613 5.43e- 1
```

```
fit4 <- update(forward1, . ~ . +hs_grad)
tidy(fit4)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  70.3       1.02       69.2  5.91e-49
## 2 murder       -0.237     0.0353     -6.72 2.18e- 8
## 3 hs_grad       0.0439    0.0161      2.72 9.09e- 3
```

```
fit5 <- update(forward1, . ~ . +frost)
tidy(fit5)
```

```
## # A tibble: 3 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  73.9       0.500      148.   2.36e-64
## 2 murder       -0.328     0.0375      -8.74 2.05e-11
## 3 frost        -0.00578   0.00266     -2.17 3.52e- 2
```

```
fit6 <- update(forward1, . ~ . +area)
tidy(fit6)
```

```
## # A tibble: 3 x 5
##   term            estimate  std.error statistic  p.value
##   <chr>              <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)  72.9        0.275        265.    2.73e-76
## 2 murder       -0.290      0.0338        -8.58  3.47e-11
## 3 area          0.00000118 0.00000146     0.806 4.24e- 1
```

```
forward2 <- update(forward1, . ~ . + hs_grad)
tidy(forward2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  70.3       1.02       69.2  5.91e-49
## 2 murder       -0.237     0.0353     -6.72 2.18e- 8
## 3 hs_grad       0.0439    0.0161      2.72 9.09e- 3
```

```
fit1 <- update(forward2, . ~ . +population)
tidy(fit1)
```

```
## # A tibble: 4 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  70.4       0.969         72.7  3.95e-49
## 2 murder       -0.266     0.0357        -7.45 1.91e- 9
## 3 hs_grad       0.0407    0.0154         2.64 1.12e- 2
## 4 population    0.0000625 0.0000259      2.41 1.99e- 2
```

```
fit2 <- update(forward2, . ~ . +income)
tidy(fit2)
```

```
## # A tibble: 4 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  70.1        1.10        64.0   1.33e-46
## 2 murder       -0.239      0.0358      -6.66  2.92e- 8
## 3 hs_grad       0.0391     0.0203       1.92  6.05e- 2
## 4 income        0.0000953  0.000239     0.398 6.92e- 1
```

```
fit3 <- update(forward2, . ~ . +illiteracy)
tidy(fit3)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  69.7       1.22      57.1   2.41e-44
## 2 murder       -0.258     0.0435    -5.93  3.63e- 7
## 3 hs_grad       0.0518    0.0188     2.76  8.25e- 3
## 4 illiteracy    0.254     0.305      0.833 4.09e- 1
```

```
fit4 <- update(forward2, . ~ . +frost)
tidy(fit4)
```

```
## # A tibble: 4 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.0       0.983       72.2  5.25e-49
## 2 murder       -0.283     0.0367      -7.71 8.04e-10
```

```
## 3 hs_grad       0.0499    0.0152       3.29 1.95e- 3
## 4 frost        -0.00691   0.00245     -2.82 6.99e- 3
```

```
fit5 <- update(forward2, . ~ . +area)
tidy(fit5)
```

```
## # A tibble: 4 x 5
##   term            estimate  std.error statistic  p.value
##   <chr>              <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)  69.9        1.16          60.1   2.30e-45
## 2 murder       -0.224      0.0404        -5.56  1.30e- 6
## 3 hs_grad       0.0504     0.0190         2.65  1.10e- 2
## 4 area         -0.00000106 0.00000162    -0.658 5.14e- 1
```

```
forward3 <- update(forward2, . ~ . + frost)
tidy(forward3)
```

```
## # A tibble: 4 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.0       0.983       72.2  5.25e-49
## 2 murder       -0.283     0.0367      -7.71 8.04e-10
## 3 hs_grad       0.0499    0.0152       3.29 1.95e- 3
## 4 frost        -0.00691   0.00245     -2.82 6.99e- 3
```

```
fit1 <- update(forward3, . ~ . +population)
tidy(fit1)
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.0       0.953         74.5  8.61e-49
## 2 murder       -0.300     0.0366        -8.20 1.77e-10
## 3 hs_grad       0.0466    0.0148         3.14 2.97e- 3
## 4 frost        -0.00594   0.00242       -2.46 1.80e- 2
## 5 population    0.0000501 0.0000251      2.00 5.20e- 2
```

```
fit2 <- update(forward3, . ~ . +income)
tidy(fit2)
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  70.8       1.05        67.4   7.53e-47
## 2 murder       -0.286     0.0373      -7.66  1.07e- 9
## 3 hs_grad       0.0436    0.0190       2.30  2.64e- 2
## 4 frost        -0.00698   0.00247     -2.83  6.96e- 3
## 5 income        0.000127  0.000223     0.571 5.71e- 1
```

```
fit3 <- update(forward3, . ~ . +illiteracy)
tidy(fit3)
```

```
## # A tibble: 5 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.5       1.32        54.2   1.28e-42
## 2 murder       -0.273     0.0411      -6.64  3.50e- 8
## 3 hs_grad       0.0450    0.0178       2.53  1.49e- 2
```

```
## 4 frost        -0.00768   0.00283    -2.72  9.36e- 3
## 5 illiteracy   -0.182     0.328      -0.554 5.82e- 1
```

```
fit4 <- update(forward3, . ~ . +area)
tidy(fit4)
```

```
## # A tibble: 5 x 5
##   term             estimate  std.error statistic  p.value
##   <chr>               <dbl>      <dbl>     <dbl>    <dbl>
## 1 (Intercept)  70.9         1.15          61.7   3.92e-45
## 2 murder       -0.279       0.0427        -6.52  5.34e- 8
## 3 hs_grad       0.0519      0.0179         2.91  5.66e- 3
## 4 frost        -0.00682     0.00251       -2.71  9.40e- 3
## 5 area         -0.000000329 0.00000154    -0.214 8.32e- 1
```

```
forward4 <- update(forward3, . ~ . + population)
tidy(forward4)
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  71.0       0.953         74.5  8.61e-49
## 2 murder       -0.300     0.0366        -8.20 1.77e-10
## 3 hs_grad       0.0466    0.0148         3.14 2.97e- 3
## 4 frost        -0.00594   0.00242       -2.46 1.80e- 2
## 5 population    0.0000501 0.0000251      2.00 5.20e- 2
```