Exercise 3.2

Apply family

useful commands: `function(){ }`, `apply()`, `tapply()`, `hist()`, `dim()`, `prod()`, `sd()`

00*. Load state data

```
?state
data(state)
# this data is stored in a slightly different way than other datasets we've used so far
states = as.data.frame(state.x77) # run this line of code to avoid later confusion
```

01*. What is the average population, income, and area of all 50 states?
Do this in one short line of code

```
apply(states, 2, mean)
```

```
## Population     Income Illiteracy   Life Exp     Murder    HS Grad      Frost
##  4246.4200  4435.8000     1.1700    70.8786     7.3780    53.1080   104.4600
##       Area
## 70735.8800
```

```
colMeans(states) # alternative
```

```
## Population     Income Illiteracy   Life Exp     Murder    HS Grad      Frost
##  4246.4200  4435.8000     1.1700    70.8786     7.3780    53.1080   104.4600
##       Area
## 70735.8800
```

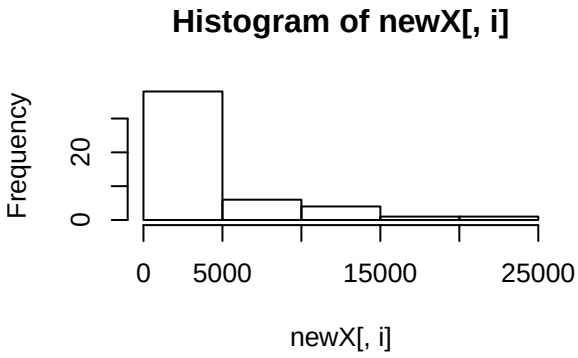
02*. What is the average area of the states from different regions of the country?
Hint: use the object state.region in your environment

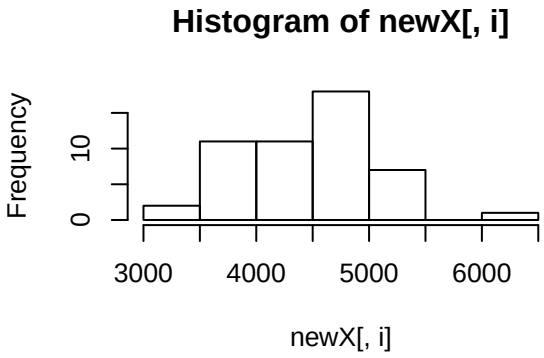
```
tapply(states$Area, state.region, mean)
```

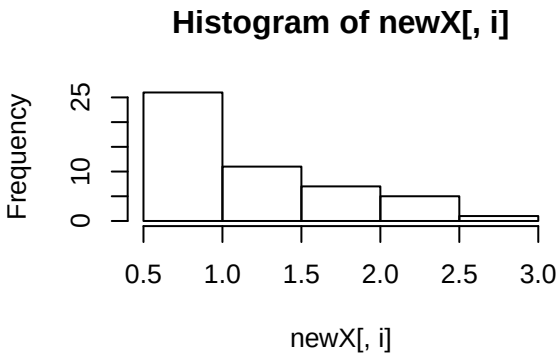
```
##     Northeast         South North Central          West
##      18141.00      54605.12      62652.00     134463.00
```

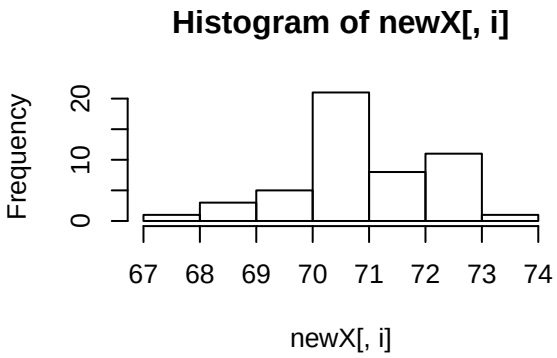
03*. Plot a histogram for each variable in the states data (Population, Income, Illiteracy etc.)

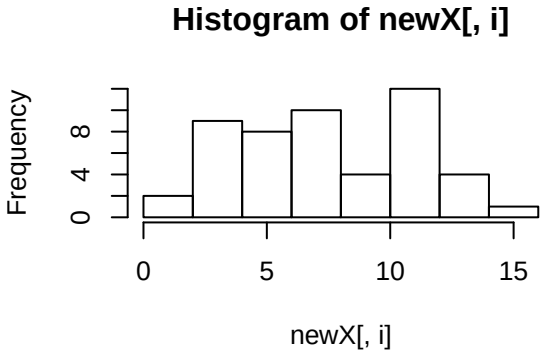
```
# 'hide' code chunk option only shows plots, no text output
dim(states)
par(mfrow = c(2,2)) # shows 2 rows and 2 columns of plots at once in a frame
apply(states, 2, hist)
```

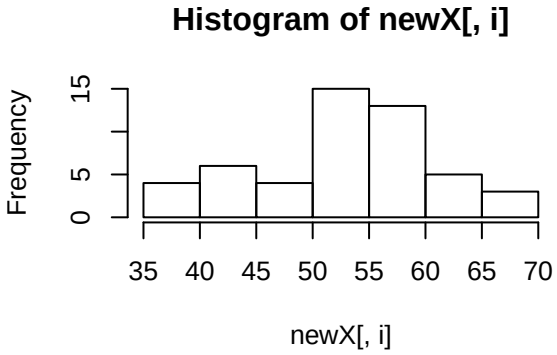


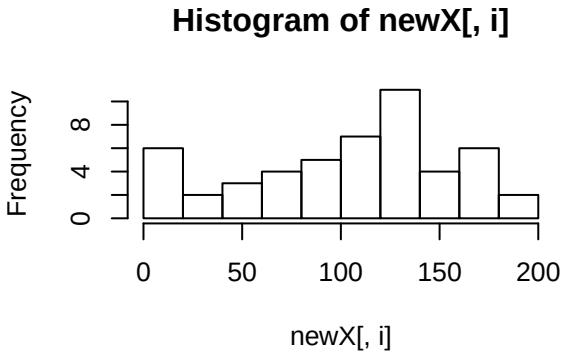


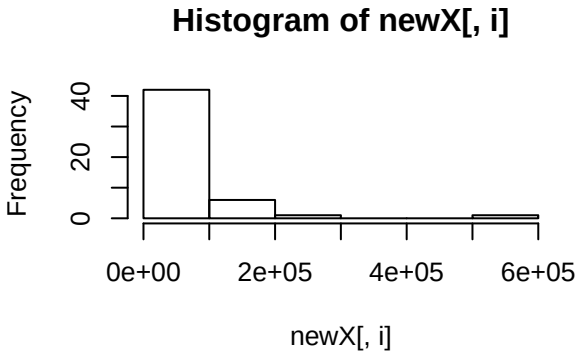












04*. Let's assume that we don't want to live in a state with high illiteracy, high murder, and many freezing days.
Also assume that each of these factors contribute equally to our opinion: (Illiteracy x Murder

x Frost) = undesirable.
Which 10 states should we avoid?
Hint: use prod(); and maybe order()

```
livability = apply(states[,c("Illiteracy", "Murder", "Frost")], # subset to variables of interest
                   1, prod) # apply product (i.e. multiplication) to each row
livability[order(livability, decreasing = T)][1:10] # top ten least livable states
```

```
##         Alaska     New Mexico South Carolina        Georgia       Kentucky
##        2576.40        2560.80        1734.20        1668.00        1611.20
## North Carolina    Mississippi      Tennessee       New York       Michigan
##        1598.40        1500.00        1309.00        1251.32        1248.75
```

BONUS 05. Standardize all the variables in the states dataset and save your answer to a new dataframe, states_standardized
Hint: to standardize a variable, you subtract the mean and divide by the standard deviation (sd)

```
states_standardized = apply(states, 2, function(x) (x-mean(x))/sd(x))
# original:
head(states)
```

```
##            Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
## Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
## Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
## Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
## Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
## California      21198   5114        1.1    71.71   10.3    62.6    20 156361
## Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
```

```
# new:
head(states_standardized)
```

```
##            Population     Income Illiteracy   Life Exp     Murder    HS Grad
## Alabama    -0.1414316 -1.3211387   1.525758 -1.3621937  2.0918101 -1.4619293
## Alaska     -0.8693980  3.0582456   0.541398 -1.1685098  1.0624293  1.6828035
## Arizona    -0.4556891  0.1533029   1.033578 -0.2447866  0.1143154  0.6180514
## Arkansas   -0.4785360 -1.7214837   1.197638 -0.1628435  0.7373617 -1.6352611
## California  3.7969790  1.1037155  -0.114842  0.6193415  0.7915396  1.1751891
## Colorado   -0.3819965  0.7294092  -0.771082  0.8800698 -0.1565742  1.3361400
##                Frost       Area
## Alabama    -1.6248292 -0.2347183
## Alaska      0.9145676  5.8093497
## Arizona    -1.7210185  0.5002047
## Arkansas   -0.7591257 -0.2202212
## California -1.6248292  1.0034903
## Colorado    1.1838976  0.3870991
```

BONUS 06. Create a histogram again for each variable in the states data, but this time label each histogram with the variable names when you plot
Hint: instead of using apply to iterate over the columns themselves, you can often iterate over the column names with sapply

```
par(mfrow = c(2,2))
sapply(colnames(states), function(x)
```

```
hist(states[ , x],
     main = x, xlab = x,
     col = "darkblue"))
```