NC STATE UNIVERSITY

Programming in R Part II

Justin Post

August 15, 2018

What do we want to be able to do?

- Restructure Data/Clean Data
- Streamline repeated sections of code
- **Improve efficiency of code**
- Write custom functions to simplify code

Efficient Code

For loops inefficient in R

- R interpreted language
- Must figure out how to evaluate code at each iteration of loop
- Slows it down

Efficient Code

For loops inefficient in R

- R interpretted language
- Must figure out how to evaluate code at each iteration of loop
- Slows it down

Vecotrized functions much faster!

- Vectorized function: works on entire vector at once
- Avoids costly computation time

Efficient Code

Some 'built-in' vectorized functions

- `colMeans()`, `rowMeans()`
- `colSums()`, `rowSums()`
- `colSds()`, `colVars()`, `colMedians()` (`matrixStats` package)
- `ifelse()`
- `apply()` family
- Create your own with `Vectorize()`

Efficient Code

- Find column means for full Batting data set
- `colMeans()` just requires a numeric data frame (array)

```
colMeans(select(Batting, G:GIDP), na.rm = TRUE)
```

```
##          G         AB          R          H        X2B        X3B
##  51.343439 141.905511  18.815544  37.139930   6.289167   1.293252
##         HR        RBI         SB         CS         BB         SO
##   2.813599  17.003975   2.976821   1.226008  13.067207  20.529712
##        IBB        HBP         SH         SF       GIDP
##   1.105870   1.056057   2.299540   1.054101   2.981018
```

Efficient Code

- Compare computational time
- `microbenchmark` package allows for easy recording of computing time

```
install.packages("microbenchmarK")
```

```
library(microbenchmark)
```

Efficient Code

- Compare computational time

```
Bat <- select(Batting, G:GIDP)
microbenchmark(
  colMeans(Bat, na.rm = TRUE)
)
```

```
## Unit: milliseconds
##                         expr      min       lq     mean   median       uq
##  colMeans(Bat, na.rm = TRUE) 6.292828 6.700414 8.966566 7.213992 8.601368
##      max neval
##  20.6466   100
```

Efficient Code

- Compare computational time

```
microbenchmark(
  for(i in 1:17){
    mean(Bat[ , i], na.rm = TRUE)
  }
)
```

```
## Unit: milliseconds
##                                                 expr      min       lq
##  for (i in 1:17) {     mean(Bat[, i], na.rm = TRUE) } 18.36648 29.49804
##      mean   median       uq      max neval
##  41.59424 40.00155 47.92906 86.60699   100
```

Efficient Code

- With vectorized functions, can easily find cool stuff
- Median number of games played for all players
- Median number of AB for players that batted
- Steps: (think `dplyr` commands!)

1. Group observations by playerID
2. Summarise variables of interest
3. Remove non numeric column
4. Coerce to matrix for use in `colMedians()`
5. Use `colMedians()` function

Efficient Code

```
library(matrixStats) #install if not installed
```

```
##
## Attaching package: 'matrixStats'
```

```
## The following object is masked from 'package:dplyr':
##
##     count
```

```
Batting %>% group_by(playerID) %>%
  summarise(totG = sum(G), totAB = sum(AB)) %>%
  select(-playerID) %>% as.matrix() %>%
  colMedians(na.rm = TRUE)
```

```
## [1] 78 71
```

- Next up, `ifelse()`

Efficient Code

- Logical statement - comparison of two quantities
  - resolves as TRUE or FALSE
- Often want to execute code logically
- logical comparison operators
  - `==`, `!=`, `>=`, `<=`, `>`, `<`
  - `&` "and"
  - `|` "or"
- logical functions
  - `is.` family (`is.numeric()`, `is.data.frame()`, etc.)

Efficient Code

If then, If then else

- Often want to execute statements conditionally
- If then else concept

```
if (condition) {
  then execute code
}

#if then else
if (condition) {
  execute this code
} else {
  execute this code
}
```

Efficient Code

If then, If then else

- Often want to execute statements conditionally
- If then else concept

```
#Or more if statements
if (condition) {
  execute this code
} else if (condition2) {
  execute this code
} else if (condition3) {
  execute this code
} else {
  #if no conditions met
  execute this code
}
```

Efficient Code

- Often create new variables

Efficient Code

- Often create new variables
- Built in data set `airquality`
    - daily air quality measurements in New York
    - from May (Day 1) to September (Day 153) in 1973

```
airquality<-tbl_df(airquality)
airquality
```

```
## # A tibble: 153 x 6
##   Ozone Solar.R  Wind  Temp Month   Day
##   <int>   <int> <dbl> <int> <int> <int>
## 1    41     190   7.4    67     5     1
## 2    36     118   8      72     5     2
## 3    12     149  12.6    74     5     3
## 4    18     313  11.5    62     5     4
## 5    NA      NA  14.3    56     5     5
## # ... with 148 more rows
```

Efficient Code

Want to code a wind category variable

- high wind days (15mph $\leq$ wind)
- windy days (10mph $\leq$ wind < 15mph)
- lightwind days (6mph $\leq$ wind < 10mph)
- calm days (wind $\leq$ 6mph)

Efficient Code

Want to code a wind category variable

- high wind days (15mph $\leq$ wind)
- windy days (10mph $\leq$ wind < 15mph)
- lightwind days (6mph $\leq$ wind < 10mph)
- calm days (wind $\leq$ 6mph)

Initial plan

- loop through each observation
- use if then else to determine wind status

Efficient Code

```
#initialize vector to save results
status<-vector()

for (i in 1:(dim(airquality)[1])){
  if(airquality$Wind[i] >= 15){
    status[i] <- "HighWind"
  } else if (airquality$Wind[i] >= 10){
    status[i] <- "Windy"
  } else if (airquality$Wind[i] >= 6){
    status[i] <- "LightWind"
  } else if (airquality$Wind[i] >= 0){
    status[i] <- "Calm"
  } else {
    status[i] <- "Error"
  }
}
```

Efficient Code

```
status
```

```
##   [1] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
##   [6] "Windy"     "LightWind" "Windy"     "HighWind"  "LightWind"
##  [11] "LightWind" "LightWind" "LightWind" "Windy"     "Windy"
##  [16] "Windy"     "Windy"     "HighWind"  "Windy"     "LightWind"
##  [21] "LightWind" "HighWind"  "LightWind" "Windy"     "HighWind"
##  [26] "Windy"     "LightWind" "Windy"     "Windy"     "Calm"
##  [31] "LightWind" "LightWind" "LightWind" "HighWind"  "LightWind"
##  [36] "LightWind" "Windy"     "LightWind" "LightWind" "Windy"
##  [41] "Windy"     "Windy"     "LightWind" "LightWind" "Windy"
##  [46] "Windy"     "Windy"     "HighWind"  "LightWind" "Windy"
##  [51] "Windy"     "LightWind" "Calm"      "Calm"      "LightWind"
##  [56] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
##  [61] "LightWind" "Calm"      "LightWind" "LightWind" "Windy"
##  [66] "Calm"      "Windy"     "Calm"      "LightWind" "Calm"
##  [71] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
##  [76] "Windy"     "LightWind" "Windy"     "LightWind" "Calm"
##  [81] "Windy"     "LightWind" "LightWind" "Windy"     "LightWind"
##  [86] "LightWind" "LightWind" "Windy"     "LightWind" "LightWind"
##  [91] "LightWind" "LightWind" "LightWind" "Windy"     "LightWind"
##  [96] "LightWind" "LightWind" "Calm"      "Calm"      "Windy"
## [101] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
```

Efficient Code

- Add it to the data set

```
airquality$status <- status
```

- Find mean temperature for each wind Status

```
airquality$status <- status
airquality %>% group_by(status) %>%
  mutate(avgTemp = mean(Temp))
```

Efficient Code

```
## # A tibble: 153 x 8
## # Groups:   status [4]
##   Ozone Solar.R  Wind  Temp Month   Day status    avgTemp
##   <int>   <int> <dbl> <int> <int> <int> <chr>       <dbl>
## 1    41     190   7.4    67     5     1 LightWind    79.4
## 2    36     118   8      72     5     2 LightWind    79.4
## 3    12     149  12.6    74     5     3 Windy        75.5
## 4    18     313  11.5    62     5     4 Windy        75.5
## 5    NA      NA  14.3    56     5     5 Windy        75.5
## # ... with 148 more rows
```

Efficient Code

- Know for loops not great
- `ifelse()` is vectorized version of `if then else`
- Syntax

```
ifelse(vector_condition, if_true_do_this, if_false_do_this)
```

Efficient Code

```
ifelse(airquality$Wind >= 15, "HighWind",
         ifelse(airquality$Wind >= 10, "Windy",
                ifelse(airquality$Wind >= 6, "LightWind", "Calm")))
```

```
##   [1] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
##   [6] "Windy"     "LightWind" "Windy"     "HighWind"  "LightWind"
##  [11] "LightWind" "LightWind" "LightWind" "Windy"     "Windy"
##  [16] "Windy"     "Windy"     "HighWind"  "Windy"     "LightWind"
##  [21] "LightWind" "HighWind"  "LightWind" "Windy"     "HighWind"
##  [26] "Windy"     "LightWind" "Windy"     "Windy"     "Calm"
##  [31] "LightWind" "LightWind" "LightWind" "HighWind"  "LightWind"
##  [36] "LightWind" "Windy"     "LightWind" "LightWind" "Windy"
##  [41] "Windy"     "Windy"     "LightWind" "LightWind" "Windy"
##  [46] "Windy"     "Windy"     "HighWind"  "LightWind" "Windy"
##  [51] "Windy"     "LightWind" "Calm"      "Calm"      "LightWind"
##  [56] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
##  [61] "LightWind" "Calm"      "LightWind" "LightWind" "Windy"
##  [66] "Calm"      "Windy"     "Calm"      "LightWind" "Calm"
##  [71] "LightWind" "LightWind" "Windy"     "Windy"     "Windy"
##  [76] "Windy"     "LightWind" "Windy"     "LightWind" "Calm"
##  [81] "Windy"     "LightWind" "LightWind" "Windy"     "LightWind"
##  [86] "LightWind" "LightWind" "Windy"     "LightWind" "LightWind"
##  [91] "LightWind" "LightWind" "LightWind" "Windy"     "LightWind"
```

Efficient Code

- Compare speed

```
loopTime<-microbenchmark(
  for (i in 1:(dim(airquality)[1])){
    if(airquality$Wind[i] >= 15){
      status[i] <- "HighWind"
    } else if (airquality$Wind[i] >= 10){
      status[i] <- "Windy"
    } else if (airquality$Wind[i] >= 6){
      status[i] <- "LightWind"
    } else if (airquality$Wind[i] >= 0){
      status[i] <- "Calm"
    } else{
      status[i] <- "Error"
    }
  }
, unit = "us")
```

Efficient Code

- Compare speed

```
vectorTime <- microbenchmark(
  ifelse(airquality$Wind >= 15, "HighWind",
           ifelse(airquality$Wind >= 10, "Windy",
                  ifelse(airquality$Wind >= 6, "LightWind", "Calm")))
, unit = "us")
```

Efficient Code (Note units!)

```
loopTime
```

```
## Unit: microseconds
##
##  for (i in 1:(dim(airquality)[1])) {     if (airquality$Wind[i] >= 15) {         status[i] <- "HighWi
##       min       lq     mean  median       uq      max neval
##  10679.76 11447.96 12500.47 11776.4 12893.96 21087.54   100
```

```
vectorTime
```

```
## Unit: microseconds
##
##  ifelse(airquality$Wind >= 15, "HighWind", ifelse(airquality$Wind >=      10, "Windy", ifelse(airqual
##      min      lq     mean  median       uq    max neval
##  142.508 143.692 149.2427 144.482 146.0615 317.78   100
```

Efficient Code

- `apply()` family of functions fast
- Check `help(apply)`
  - We'll look at `apply()`, `sapply()`, `lapply()`

Efficient Code

- `apply()` family of functions fast
- Check `help(apply)`
  - We'll look at `apply()`, `sapply()`, `lapply()`
    - Use `apply()` to find summary for columns of airquality data

```
apply(X = select(airquality, Ozone:Temp), MARGIN = 2,
      FUN = summary, na.rm = TRUE)
```

Efficient Code

- Keeps data numeric, keeps labels!

```
## $Ozone
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   18.00   31.50   42.13   63.25  168.00      37
##
## $Solar.R
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     7.0   115.8   205.0   185.9   258.8   334.0       7
##
## $Wind
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.700   7.400   9.700   9.958  11.500  20.700
##
## $Temp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   72.00   79.00   77.88   85.00   97.00
```

Efficient Code

- Use `lapply()` to apply function to lists
- Obtain a list object

```
fit <- lm(Ozone ~ Wind, data = airquality)
fit <- list(fit$residuals, fit$effects, fit$fitted.values)
```

Efficient Code

```
fit[[1]]
```

```
##           1           2           3           4           6           7
## -14.7960653 -16.4655116 -14.9312663 -15.0372815  13.8358563 -26.1349578
##           8           9          11          12          13          14
##  -1.2701589  22.7006553 -51.5715267 -27.0289427 -34.8044041 -22.3678352
##          15          16          17          18          19          20
##  -5.6007126 -19.0372815   3.7381799  11.2640864  -3.0372815 -32.0289427
##          21          22          23          24          28          29
## -42.0289427   6.2724252 -39.0289427   1.7381799  -7.2618201  30.8358563
##          30          31          38          40          41          44
##  49.7673658 -18.7960653 -14.0289427  50.7298411   5.9627185 -29.4655116
##          47          48          49          50          51          62
##   6.8358563  55.0312090 -25.8044041 -21.0372815 -26.6983889  60.8858892
##          63          64          66          67          68          69
##   3.1955959 -13.8044041  -7.3386494   3.6321648   8.4368121  35.0979195
##          70          71          73          74          76          77
##  31.7673658  29.2039347  -7.4946974  12.8358563 -10.4946974 -10.5715267
##          78          79          80          81          82          85
##  -4.6983889  -0.9020805  10.4368121  29.9627185 -42.5715267  30.8650422
##          86          87          88          89          90          91
##  55.5344884 -29.1349578  21.7381799  26.2039347  -5.7960653   8.2039347
##          92          93          94          95          96          97
```

Efficient Code

```
fit[[2]]
```

```
##  (Intercept)         Wind
## -453.7465588 -212.8004841  -14.7950343  -14.4827016   13.0973609
##
##  -24.4774608   -1.5903064   19.9845153  -49.2674921  -25.7897935
##
##  -33.3750969  -21.5850656   -5.6926704  -18.4827016    4.1026017
##
##    9.1944841   -2.4827016  -30.7897935  -40.7897935    4.8873922
##
##  -37.7897935    2.1026017   -6.8973983   30.0973609   52.5277800
##
##  -16.6821888  -12.7897935   50.4096936    6.5172984  -27.5798248
##
##    6.0973609   52.0868793  -24.3750969  -20.4827016  -25.6874296
##
##   64.2548094    4.6249031  -12.3750969   -4.1598873    4.4149344
##
##   11.4254160   37.6301439   34.5277800   31.3178112   -8.0050031
##
##   12.0973609  -11.0050031   -8.2674921   -3.6874296    1.6301439
##
```

Efficient Code

- Apply `mean()` function to each list element

```
lapply(X = fit, FUN = mean)
```

```
## [[1]]
## [1] -5.731915e-16
##
## [[2]]
## [1] -4.333566
##
## [[3]]
## [1] 42.12931
```

Efficient Code

- Use `sapply()` similar but returns a vector if possible

```
sapply(X = fit, FUN = mean)
```

```
## [1] -5.731915e-16 -4.333566e+00  4.212931e+01
```

- `apply()` functions not as good as `colMeans()` type functions

```
air2 <- select(airquality, Ozone:Day)
microbenchmark(apply(X = air2, MARGIN = 2, FUN = mean, na.rm = TRUE))
```

```
## Unit: microseconds
##                                                 expr     min      lq
##  apply(X = air2, MARGIN = 2, FUN = mean, na.rm = TRUE) 149.219 153.364
##      mean  median       uq     max neval
##  167.4528 156.324 169.3515 345.414   100
```

```
microbenchmark(colMeans(air2, na.rm = TRUE))
```

```
## Unit: microseconds
##                          expr  min      lq     mean  median       uq
##  colMeans(air2, na.rm = TRUE) 75.4 77.5705 106.2733 79.9395 149.6135
##      max neval
##  286.595   100
```

Recap!

- Vectorized functions fast!
- 'Built-in' vectorized functions
  - `colMeans()`, `rowMeans()`
  - `colSums()`, `rowSums()`
  - `colSds()`, `colVars()`, `colMedians()` (`matrixStats` package)
  - `ifelse()`
  - `apply()` family

Activity

- **Vectorized Functions Activity** instructions available on web
- Work in small groups
- Ask questions! TAs and I will float about the room
- Feel free to ask questions about anything you didn't understand as well!

What do we want to be able to do?

- Restructure Data/Clean Data
- Streamline repeated sections of code
- Improve efficiency of code
- **Write custom functions to simplify code**

Writing Functions

- Knowing how to write **functions** vital to custom analyses!
- Function writing syntax

```
nameOfFunction <- function(input1, input2, ...) {
  #code
  #return something with return()
  #or returns last value
}
```

Writing Functions

- Can look at code for functions

```
var
```

```
## function (x, y = NULL, na.rm = FALSE, use)
## {
##     if (missing(use))
##         use <- if (na.rm)
##             "na.or.complete"
##         else "everything"
##     na.method <- pmatch(use, c("all.obs", "complete.obs", "pairwise.complete.obs",
##         "everything", "na.or.complete"))
##     if (is.na(na.method))
##         stop("invalid 'use' argument")
##     if (is.data.frame(x))
##         x <- as.matrix(x)
##     else stopifnot(is.atomic(x))
##     if (is.data.frame(y))
##         y <- as.matrix(y)
##     else stopifnot(is.atomic(y))
##     .Call(C_cov, x, y, na.method, FALSE)
## }
```

Writing Functions

- Can look at code for functions

```
colMeans
```

```
## function (x, na.rm = FALSE, dims = 1L)
## {
##     if (is.data.frame(x))
##         x <- as.matrix(x)
##     if (!is.array(x) || length(dn <- dim(x)) < 2L)
##         stop("'x' must be an array of at least two dimensions")
##     if (dims < 1L || dims > length(dn) - 1L)
##         stop("invalid 'dims'")
##     n <- prod(dn[id <- seq_len(dims)])
##     dn <- dn[-id]
##     z <- if (is.complex(x))
##         .Internal(colMeans(Re(x), n, prod(dn), na.rm)) + (0+1i) *
##             .Internal(colMeans(Im(x), n, prod(dn), na.rm))
##     else .Internal(colMeans(x, n, prod(dn), na.rm))
##     if (length(dn) > 1L) {
##         dim(z) <- dn
##         dimnames(z) <- dimnames(x)[-id]
##     }
##     else names(z) <- dimnames(x)[[dims + 1L]]
```

Writing Functions

- Can look at code for functions

```
mean
```

```
## function (x, ...)
## UseMethod("mean")
## <bytecode: 0x000000001d6e78f8>
## <environment: namespace:base>
```

Writing Functions

- Can look at code for functions

```
mean.default
```

```
## function (x, trim = 0, na.rm = FALSE, ...)
## {
##     if (!is.numeric(x) && !is.complex(x) && !is.logical(x)) {
##         warning("argument is not numeric or logical: returning NA")
##         return(NA_real_)
##     }
##     if (na.rm)
##         x <- x[!is.na(x)]
##     if (!is.numeric(trim) || length(trim) != 1L)
##         stop("'trim' must be numeric of length one")
##     n <- length(x)
##     if (trim > 0 && n) {
##         if (is.complex(x))
##             stop("trimmed means are not defined for complex data")
##         if (anyNA(x))
##             return(NA_real_)
##         if (trim >= 0.5)
##             return(stats::median(x, na.rm = FALSE))
##         lo <- floor(n * trim) + 1
```

Writing Functions

- Goal: Create a `standardize()` function
- Take vector of values
  - subtract mean
  - divide by standard deviation
- z-score idea
- Formula: For value i,

  (value[i] - mean(value)) / sd(value)

Writing Functions

```
nameOfFunction <- function(input1, input2, ...) {
  #code
  #return something with return()
  #or returns last value
}

standardize <- function(vector) {
  return((vector - mean(vector)) / sd(vector))
}
```

Writing Functions

- Now use it!

```
data <- runif(5)
```

```
data
```

```
## [1] 0.4547721 0.9785136 0.4703607 0.3316216 0.2029009
```

```
result <- standardize(data)
```

```
result
```

```
## [1] -0.1114508  1.6648273 -0.0585820 -0.5291180 -0.9656765
```

Writing Functions

- Check result has mean 0 and sd 1

```
mean(result)
```

```
## [1] 7.075503e-17
```

```
sd(result)
```

```
## [1] 1
```

Writing Functions

- Goal: Add more inputs
- Make centering optional
- Make scaling optional

```
standardize <- function(vector, center, scale) {
  if (center == TRUE) {
    vector <- vector - mean(vector)
  }
  if (scale == TRUE) {
    vector <- vector / sd(vector)
  }
  return(vector)
}
```

Writing Functions

```
result <- standardize(data, center = TRUE, scale = TRUE)
result
```

```
## [1] -0.1114508  1.6648273 -0.0585820 -0.5291180 -0.9656765
```

```
result <- standardize(data, center = FALSE, scale = TRUE)
result
```

```
## [1] 1.5423674 3.3186455 1.5952362 1.1247002 0.6881418
```

Writing Functions

- Give center and scale default arguments

```
standardize <- function(vector, center = TRUE, scale = TRUE) {
  #center and scale if appropriate
  if (center == TRUE) {
    vector <- vector - mean(vector)
  }
  if (scale == TRUE) {
    vector <- vector / sd(vector)
  }
  return(vector)
}
```

Writing Functions

```
result <- standardize(data, center = TRUE, scale = TRUE)
result
```

```
## [1] -0.1114508  1.6648273 -0.0585820 -0.5291180 -0.9656765
```

```
#same call
result <- standardize(data)
result
```

```
## [1] -0.1114508  1.6648273 -0.0585820 -0.5291180 -0.9656765
```

Writing Functions

- Return more than 1 object by returning a list
- Goal: Also return
  - `mean()` of original data
  - `sd()` of original data

Writing Functions

```
standardize <- function(vector, center = TRUE, scale = TRUE) {
  #get attributes to return
  mean <- mean(vector)
  stdev <- sd(vector)
  #center and scale if appropriate
  if (center == TRUE) {
    vector <- vector - mean
  }
  if (scale == TRUE) {
    vector <- vector / stdev
  }
  #return a list of objects
  return(list(vector, mean, stdev))
}
```

Writing Function

```
result <- standardize(data)
result
```

```
## [[1]]
## [1] -0.1114508  1.6648273 -0.0585820 -0.5291180 -0.9656765
##
## [[2]]
## [1] 0.4876338
##
## [[3]]
## [1] 0.2948533
```

```
result[[2]]
```

```
## [1] 0.4876338
```

Writing Functions

- Fancy up what we return by giving names

```
standardize <- function(vector, center = TRUE, scale = TRUE) {
  #get attributes to return
  mean <- mean(vector)
  stdev <- sd(vector)
  #center and scale if appropriate
  if (center == TRUE) {
    vector <- vector - mean
  }
  if (scale == TRUE) {
    vector <- vector / stdev
  }
  #return a list of objects
  return(list(result = vector, mean = mean, sd = stdev))
}
```

Writing Functions

```
result <- standardize(data, center = TRUE, scale = TRUE)
result
```

```
## $result
## [1] -0.1114508  1.6648273 -0.0585820 -0.5291180 -0.9656765
##
## $mean
## [1] 0.4876338
##
## $sd
## [1] 0.2948533
```

```
result$sd
```

```
## [1] 0.2948533
```

Writing Functions

- Can bring in unnamed arguments
- Arguments that can be used by functions **inside** your function
- Done already in `apply()`

apply

```
## function (X, MARGIN, FUN, ...)
## {
##     FUN <- match.fun(FUN)
##     dl <- length(dim(X))
##     if (!dl)
##         stop("dim(X) must have a positive length")
##     if (is.object(X))
##         X <- if (dl == 2L)
##             as.matrix(X)
##         else as.array(X)
##     d <- dim(X)
##     dn <- dimnames(X)
##     ds <- seq_len(dl)
##     if (is.character(MARGIN)) {
##         if (is.null(dnn <- names(dn)))
##             stop("'X' must have named dimnames")
```

Writing Functions

```
apply(X = select(airquality, Ozone:Temp), MARGIN = 2,
      FUN = summary, na.rm = TRUE)
```

```
## $Ozone
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00   18.00   31.50   42.13   63.25  168.00      37
##
## $Solar.R
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     7.0   115.8   205.0   185.9   258.8   334.0       7
##
## $Wind
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.700   7.400   9.700   9.958  11.500  20.700
##
## $Temp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   72.00   79.00   77.88   85.00   97.00
```

Writing Functions

- Add unnamed arguments to our function

```
standardize <- function(vector, center = TRUE, scale = TRUE, ...) {
  #get attributes to return
  mean <- mean(vector, ...)
  stdev <- sd(vector, ...)
  #center and scale if appropriate
  if (center == TRUE) {
    vector <- vector - mean
  }
  if (scale == TRUE) {
    vector <- vector / stdev
  }
  #return a list of objects
  return(list(result = vector, mean = mean, sd = stdev))
}
```

Writing Functions

```
sData <- standardize(airquality$Ozone, na.rm = TRUE)
sData$mean
```

```
## [1] 42.12931
```

```
sData$sd
```

```
## [1] 32.98788
```

```
sData$result
```

```
##   [1] -0.03423409 -0.18580489 -0.91334473 -0.73145977          NA
##   [6] -0.42831817 -0.57988897 -0.70114561 -1.03460136          NA
##  [11] -1.06491552 -0.79208809 -0.94365889 -0.85271641 -0.73145977
##  [16] -0.85271641 -0.24643321 -1.09522968 -0.36768985 -0.94365889
##  [21] -1.24680048 -0.94365889 -1.15585800 -0.30706153          NA
##  [26]          NA          NA -0.57988897  0.08702254  2.20901373
##  [31] -0.15549073          NA          NA          NA          NA
##  [36]          NA          NA -0.39800401          NA  0.87519070
##  [41] -0.09486241          NA          NA -0.57988897          NA
##  [46]          NA -0.64051729 -0.15549073 -0.67083145 -0.91334473
```

Recap!

- Function writing opens R up!
- Syntax

```
nameOfFunction <- function(input1, input2, ...) {
  #code
  #return something with return()
  #or returns last value
}
```

- Can set defaults in function definition
- Can return a named list
- Can give unnamed arguments for use

What do we want to be able to do?

- Restructure Data/Clean Data
- Streamline repeated sections of code
- Improve efficiency of code
- Write custom functions to simplify code

Activity

- **Function Writing Activity** instructions available on web
- Work in small groups
- Ask questions! TAs and I will float about the room
- Feel free to ask questions about anything you didn't understand as well!
- Thanks for coming!