R intro1

Module 1

8/25/2018

R has many built-in datasets. This one is Old Faithful at Yellowstone National Park. eruptions and waiting are the variables (sometimes called duration and interval)

```
data(faithful)
head(faithful)
```

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

Variables objects inside the faith dataset so to access just one variable, you need to either use a two-level name (`datasetname$variablename`) such as `faith$eruptions`. The other option is to use `attach()` to access the variables without the two-level name. Make sure to use `detach()` at the end of your session.

```
# eruptions (will give an error)
faithful$eruptions
```

```
  [1] 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350 1.833
 [12] 3.917 4.200 1.750 4.700 2.167 1.750 4.800 1.600 4.250 1.800 1.750
 [23] 3.450 3.067 4.533 3.600 1.967 4.083 3.850 4.433 4.300 4.467 3.367
 [34] 4.033 3.833 2.017 1.867 4.833 1.833 4.783 4.350 1.883 4.567 1.750
 [45] 4.533 3.317 3.833 2.100 4.633 2.000 4.800 4.716 1.833 4.833 1.733
 [56] 4.883 3.717 1.667 4.567 4.317 2.233 4.500 1.750 4.800 1.817 4.400
 [67] 4.167 4.700 2.067 4.700 4.033 1.967 4.500 4.000 1.983 5.067 2.017
 [78] 4.567 3.883 3.600 4.133 4.333 4.100 2.633 4.067 4.933 3.950 4.517
 [89] 2.167 4.000 2.200 4.333 1.867 4.817 1.833 4.300 4.667 3.750 1.867
[100] 4.900 2.483 4.367 2.100 4.500 4.050 1.867 4.700 1.783 4.850 3.683
[111] 4.733 2.300 4.900 4.417 1.700 4.633 2.317 4.600 1.817 4.417 2.617
[122] 4.067 4.250 1.967 4.600 3.767 1.917 4.500 2.267 4.650 1.867 4.167
[133] 2.800 4.333 1.833 4.383 1.883 4.933 2.033 3.733 4.233 2.233 4.533
[144] 4.817 4.333 1.983 4.633 2.017 5.100 1.800 5.033 4.000 2.400 4.600
[155] 3.567 4.000 4.500 4.083 1.800 3.967 2.200 4.150 2.000 3.833 3.500
[166] 4.583 2.367 5.000 1.933 4.617 1.917 2.083 4.583 3.333 4.167 4.333
[177] 4.500 2.417 4.000 4.167 1.883 4.583 4.250 3.767 2.033 4.433 4.083
[188] 1.833 4.417 2.183 4.800 1.833 4.800 4.100 3.966 4.233 3.500 4.366
[199] 2.250 4.667 2.100 4.350 4.133 1.867 4.600 1.783 4.367 3.850 1.933
[210] 4.500 2.383 4.700 1.867 3.833 3.417 4.233 2.400 4.800 2.000 4.150
[221] 1.867 4.267 1.750 4.483 4.000 4.117 4.083 4.267 3.917 4.550 4.083
[232] 2.417 4.183 2.217 4.450 1.883 1.850 4.283 3.950 2.333 4.150 2.350
[243] 4.933 2.900 4.583 3.833 2.083 4.367 2.133 4.350 2.200 4.450 3.567
[254] 4.500 4.150 3.817 3.917 4.450 2.000 4.283 4.767 4.533 1.850 4.250
[265] 1.983 2.250 4.750 4.117 2.150 4.417 1.817 4.467
```

```
attach(faithful)
eruptions
```

```
  [1] 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350 1.833
 [12] 3.917 4.200 1.750 4.700 2.167 1.750 4.800 1.600 4.250 1.800 1.750
 [23] 3.450 3.067 4.533 3.600 1.967 4.083 3.850 4.433 4.300 4.467 3.367
 [34] 4.033 3.833 2.017 1.867 4.833 1.833 4.783 4.350 1.883 4.567 1.750
 [45] 4.533 3.317 3.833 2.100 4.633 2.000 4.800 4.716 1.833 4.833 1.733
 [56] 4.883 3.717 1.667 4.567 4.317 2.233 4.500 1.750 4.800 1.817 4.400
 [67] 4.167 4.700 2.067 4.700 4.033 1.967 4.500 4.000 1.983 5.067 2.017
 [78] 4.567 3.883 3.600 4.133 4.333 4.100 2.633 4.067 4.933 3.950 4.517
 [89] 2.167 4.000 2.200 4.333 1.867 4.817 1.833 4.300 4.667 3.750 1.867
[100] 4.900 2.483 4.367 2.100 4.500 4.050 1.867 4.700 1.783 4.850 3.683
[111] 4.733 2.300 4.900 4.417 1.700 4.633 2.317 4.600 1.817 4.417 2.617
[122] 4.067 4.250 1.967 4.600 3.767 1.917 4.500 2.267 4.650 1.867 4.167
[133] 2.800 4.333 1.833 4.383 1.883 4.933 2.033 3.733 4.233 2.233 4.533
[144] 4.817 4.333 1.983 4.633 2.017 5.100 1.800 5.033 4.000 2.400 4.600
[155] 3.567 4.000 4.500 4.083 1.800 3.967 2.200 4.150 2.000 3.833 3.500
[166] 4.583 2.367 5.000 1.933 4.617 1.917 2.083 4.583 3.333 4.167 4.333
[177] 4.500 2.417 4.000 4.167 1.883 4.583 4.250 3.767 2.033 4.433 4.083
[188] 1.833 4.417 2.183 4.800 1.833 4.800 4.100 3.966 4.233 3.500 4.366
[199] 2.250 4.667 2.100 4.350 4.133 1.867 4.600 1.783 4.367 3.850 1.933
[210] 4.500 2.383 4.700 1.867 3.833 3.417 4.233 2.400 4.800 2.000 4.150
[221] 1.867 4.267 1.750 4.483 4.000 4.117 4.083 4.267 3.917 4.550 4.083
[232] 2.417 4.183 2.217 4.450 1.883 1.850 4.283 3.950 2.333 4.150 2.350
[243] 4.933 2.900 4.583 3.833 2.083 4.367 2.133 4.350 2.200 4.450 3.567
[254] 4.500 4.150 3.817 3.917 4.450 2.000 4.283 4.767 4.533 1.850 4.250
[265] 1.983 2.250 4.750 4.117 2.150 4.417 1.817 4.467
```

Reading in an external cvs (comma separated values) file:

```
# faith=read.csv('S:/Courses/stat-renaes/faithdata.csv',header=T)
```

Or use the "Import dataset" option in the Environment window (upper right window in RStudio)

To see the first 6 observations, use `head()`

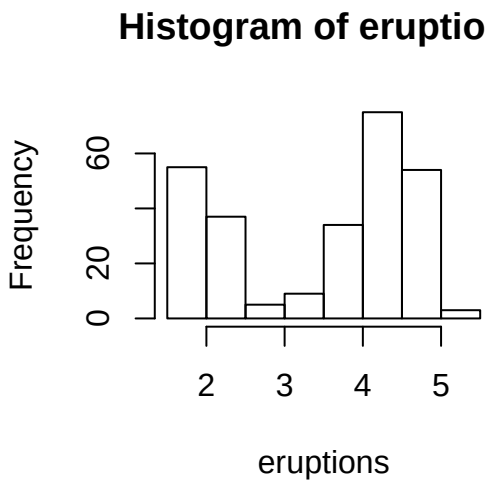
```
head(faithful)
```

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

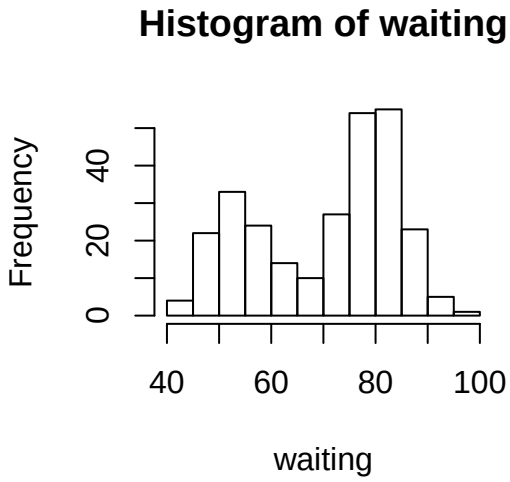
You can also just type in the dataset name in the console to see the whole dataset. Another option to that is to use `View()` (example: `View(faithful)`

Now to make some graphs. Histograms use `hist()`, boxplots use `boxplot()` and scatterplots use `plot()`.

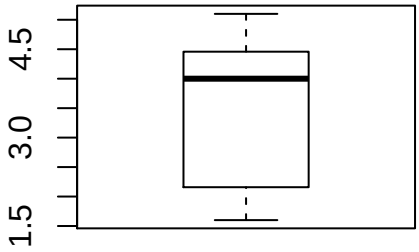
```
hist(eruptions)
```




```
hist(waiting)
```

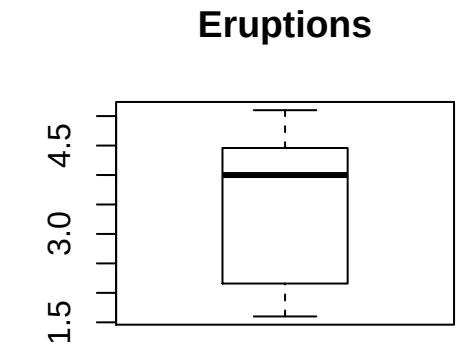



```
boxplot(eruptions)
```

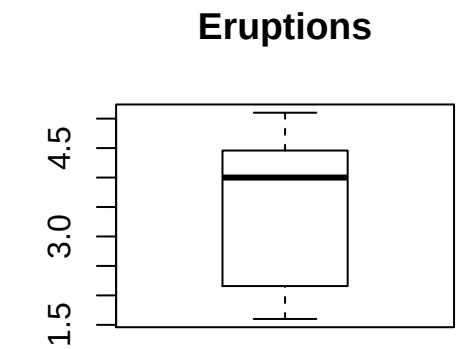


However, boxplots don't default with a title or axis labels. Use the `main=''` option in the function or use the command `title()` after the plot function. `xlab=''` and `ylab=''` are for the axis labels.

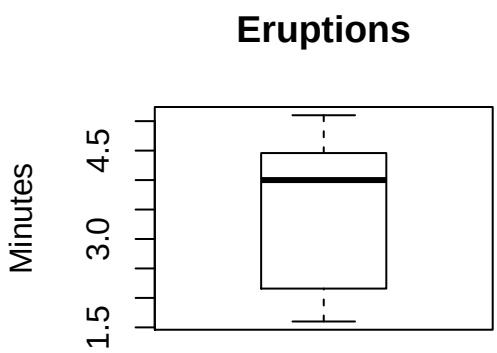
```r
boxplot(eruptions,main='Eruptions')
```



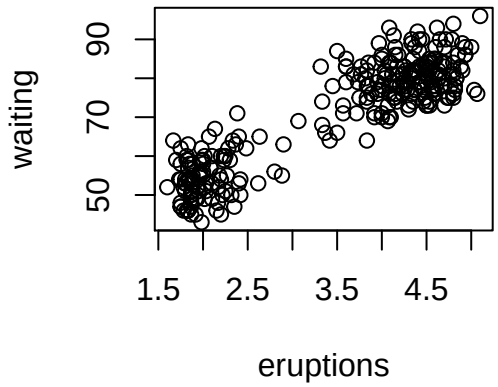

```r
# or
boxplot(eruptions); title('Eruptions')
```




```r
# also axis labels
boxplot(eruptions,xlab='Eruption duration',ylab='Minutes'); title('Eruptions')
```

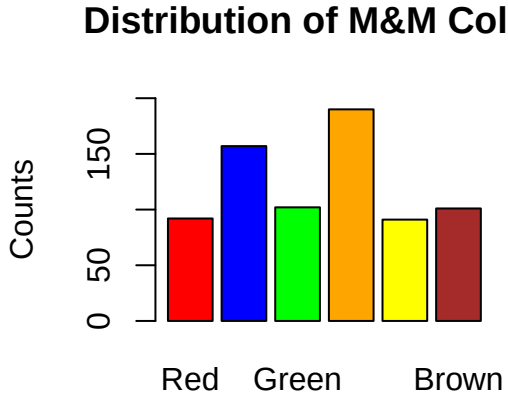



```r
# scatterplot x=eruptions y=waiting
plot(eruptions, waiting)
```



Bar graphs are helpful in certain situations (not with the `faithful` data).

```
colors=c('Red','Blue','Green','Orange','Yellow','Brown')
observed=c(92,157,102,190,91,101)
barplot(observed,names.arg=colors,col=colors,ylim=c(0,200),
  ylab='Counts',main="Distribution of M&M Colors")
```



Summary statistics can be calculated many (many!) different ways. Here are jsut a few:

Individual commands:

```
mean(eruptions); mean(waiting)
```

```
[1] 3.487783
```

```
[1] 70.89706
```

```
var(eruptions); var(waiting)
```

```
[1] 1.302728
```

```
[1] 184.8233
```

```
sd(eruptions); sd(waiting)
```

```
[1] 1.141371
```

```
[1] 13.59497
```

```
median(eruptions); median(waiting)
```

```
[1] 4
```

```
[1] 76
```

```
max(eruptions); max(waiting)
```

```
[1] 5.1
```

```
[1] 96
```

```
min(eruptions); min(waiting)
```

```
[1] 1.6
```

```
[1] 43
```

```
length(eruptions) # gives the sample size
```

```
[1] 272
```

Using summary() but only gives mean, min, max, median, q1, q3

```
summary(eruptions); summary(waiting)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.600   2.163   4.000   3.488   4.454   5.100
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   43.0    58.0    76.0    70.9    82.0    96.0
```

```
# or use summary(eruptions); summary(waiting)
summary(faithful)
```

```
   eruptions        waiting
 Min.   :1.600   Min.   :43.0
 1st Qu.:2.163   1st Qu.:58.0
 Median :4.000   Median :76.0
 Mean   :3.488   Mean   :70.9
 3rd Qu.:4.454   3rd Qu.:82.0
 Max.   :5.100   Max.   :96.0
```

Using another command but it is not in the base packages of commands we have. We will need to use a package in R. Packages install functions that are not in the base version. This one is called the stargazer package.

Installing a package. For this one, there is a `#` in front of the command because I have installed this package before and do not need to reinstall. You will need to remove the `#` before `install.packages()` to actually run and install. Unless you are prompted to, never reinstall packages. Once a package is installed, you need to load the package using the `library()` command.

```
# install.packages("stargazer")
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

 R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

Using `stargazer()` needs to have an option `type='text'` or it will output what may appear to some as jibberish, but it will output LaTeX code, but no helpful (or nice-looking) output.

```
stargazer(faithful,type='text')
```

```
=========================================
Statistic  N   Mean  St. Dev.  Min   Max
-----------------------------------------
eruptions 272 3.488   1.141   1.600 5.100
waiting   272 70.897  13.595   43    96
-----------------------------------------
```

Here is one more. It is the `stat.desc()` command in the `pastecs` package (run `install.packages('pastecs')` to install it for the first time)

```
library(pastecs)
stat.desc(faithful)
```

```
              eruptions      waiting
nbr.val      272.0000000 2.720000e+02
nbr.null       0.0000000 0.000000e+00
nbr.na         0.0000000 0.000000e+00
min            1.6000000 4.300000e+01
max            5.1000000 9.600000e+01
range          3.5000000 5.300000e+01
sum          948.6770000 1.928400e+04
median         4.0000000 7.600000e+01
mean           3.4877831 7.089706e+01
SE.mean        0.0692058 8.243164e-01
CI.mean.0.95   0.1362494 1.622878e+00
var            1.3027283 1.848233e+02
std.dev        1.1413713 1.359497e+01
coef.var       0.3272483 1.917565e-01
```

Finding the mode: there is no command for mode but here is some code that will work: if there is no mode, it will list all observations with a `1` under the values.

```
erup=table(eruptions)
erup[max(erup)==erup]
```

```
eruptions
1.867   4.5
    8     8
```

```
wait=table(waiting)
wait[max(wait)==wait]
```

```
78
15
```