R Cheatsheet

Copyright 2014, Bruce Beauchamp

Contents

[R Cheatsheet 1](#_Toc466388815)

[Best Subset Regression 4](#_Toc466388816)

[Control statements 4](#_Toc466388817)

[for-next loops 4](#_Toc466388818)

[Iterating through a vector 4](#_Toc466388819)

[data.frames 5](#_Toc466388820)

[Aggregate a column by another column 5](#_Toc466388821)

[Aggregate multiple columns 5](#_Toc466388822)

[Append a column to a data.frame 6](#_Toc466388823)

[Append a row to a data.frame 6](#_Toc466388824)

[Browse a data.frame 7](#_Toc466388825)

[Convert a column to date data type 7](#_Toc466388826)

[Create a data.frame from scratch 7](#_Toc466388827)

[Create a factor column based upon values in an existing column 8](#_Toc466388828)

[Edit a data.frame in table form 8](#_Toc466388829)

[Expose the fields in a data.frame to the development environment 8](#_Toc466388830)

[Get a cell of data from a data.frame 8](#_Toc466388831)

[Get the column names of a data.frame 9](#_Toc466388832)

[Get the number of rows in a data.frame 9](#_Toc466388833)

[Get a Row of Data from a data.frame 9](#_Toc466388834)

[Import a data.frame column as a formatted date 9](#_Toc466388835)

[Look at the first part of a dataframe 9](#_Toc466388836)

[Look at the structure of a data.frame 9](#_Toc466388837)

[Remove rows containing NA values from a data.frame 10](#_Toc466388838)

[Select Rows which meet criteria 10](#_Toc466388839)

[Select Rows which meet criteria using sqldf 11](#_Toc466388840)

[Select Columns from a data.frame by index 11](#_Toc466388841)

[Subsetting rows selecting on more than one value 11](#_Toc466388842)

[Data Types 12](#_Toc466388843)

[Get the type of an object 12](#_Toc466388844)

[Dates and Times 12](#_Toc466388845)

[Convert a String to a Date 12](#_Toc466388846)

[Distributions 12](#_Toc466388847)

[Beta 12](#_Toc466388848)

[Cumulative Probability for a Beta Distribution 12](#_Toc466388849)

[Environment 12](#_Toc466388850)

[Change the Working Directory 12](#_Toc466388851)

[Get the Working Directory 12](#_Toc466388852)

[List the Files in the Working Directory 12](#_Toc466388853)

[Load a Function into the Environment 13](#_Toc466388854)

[List the Functions Available in the Environment 13](#_Toc466388855)

[Files and Directories 13](#_Toc466388856)

[Create a directory 13](#_Toc466388857)

[Download a file from the web 13](#_Toc466388858)

[See if a directory or file exists 13](#_Toc466388859)

[Libraries 13](#_Toc466388860)

[Loading Libraries 13](#_Toc466388861)

[Linear Regression 14](#_Toc466388862)

[Confidence interval for a regression line fit 14](#_Toc466388863)

[Generate the Regression line 14](#_Toc466388864)

[Get more information about the linear regression 14](#_Toc466388865)

[Plot a regression line on a plot 14](#_Toc466388866)

[Matrices 15](#_Toc466388867)

[Create a matrix 15](#_Toc466388868)

[Misc 15](#_Toc466388869)

[Printing a variable to the console 15](#_Toc466388870)

[Packages 15](#_Toc466388871)

[Installing packages 15](#_Toc466388872)

[Plotting data 16](#_Toc466388873)

[Adding a legend to a base plot 16](#_Toc466388874)

[Bar Plots 16](#_Toc466388875)

[faceted plots (ggplot2) 17](#_Toc466388876)

[Line Plots 17](#_Toc466388877)

[Histograms 18](#_Toc466388878)

[Line plots with multiple lines 18](#_Toc466388879)

[Multiple plots in a grid 19](#_Toc466388880)

[Multiple plots in quadrants 20](#_Toc466388881)

[Plotting to a PNG file 20](#_Toc466388882)

[Scatter plot of y vs x 21](#_Toc466388883)

[Using Base Plot 21](#_Toc466388884)

[Reading Data 22](#_Toc466388885)

[Read a CSV file into a data.frame 22](#_Toc466388886)

[Read from an excel spreadsheet 22](#_Toc466388887)

[Statistics 22](#_Toc466388888)

[p-value for a Binomial Distribution 22](#_Toc466388889)

[Strings 22](#_Toc466388890)

[Concatenate two strings 22](#_Toc466388891)

[Padding a string with zeros 22](#_Toc466388892)

[Vectors 23](#_Toc466388893)

[Concatenate two string vectors 23](#_Toc466388894)

[Creating a vector 23](#_Toc466388895)

[Iterating over a vector 23](#_Toc466388896)

[Length of a vector 23](#_Toc466388897)

[Return valid elements of a vector 23](#_Toc466388898)

# Best Subset Regression

Use the leaps() package.

library(leaps)

regfit.full = regsubsets(Salary~., data=Hitters)

# Control statements

## for-next loops

for (i in 1:50 ) {

foo.squared[i] = foo[i]^2

}

## Iterating through a vector

id = 1:3

for (i in id) {

cat(i, "\n")

}

|  |
| --- |
| 1  2  3 |
|  |
| |  | | --- | |  | |

# data.frames

## Aggregate a column by another column

> head(NEI)

fips SCC Pollutant Emissions type year

4 09001 10100401 PM25-PRI 15.714 POINT 1999

8 09001 10100404 PM25-PRI 234.178 POINT 1999

12 09001 10100501 PM25-PRI 0.128 POINT 1999

16 09001 10200401 PM25-PRI 2.036 POINT 1999

20 09001 10200504 PM25-PRI 0.388 POINT 1999

24 09001 10200602 PM25-PRI 1.490 POINT 1999

> test <- aggregate(Emissions~year, NEI, sum)

> class(test)

[1] "data.frame"

> head(test)

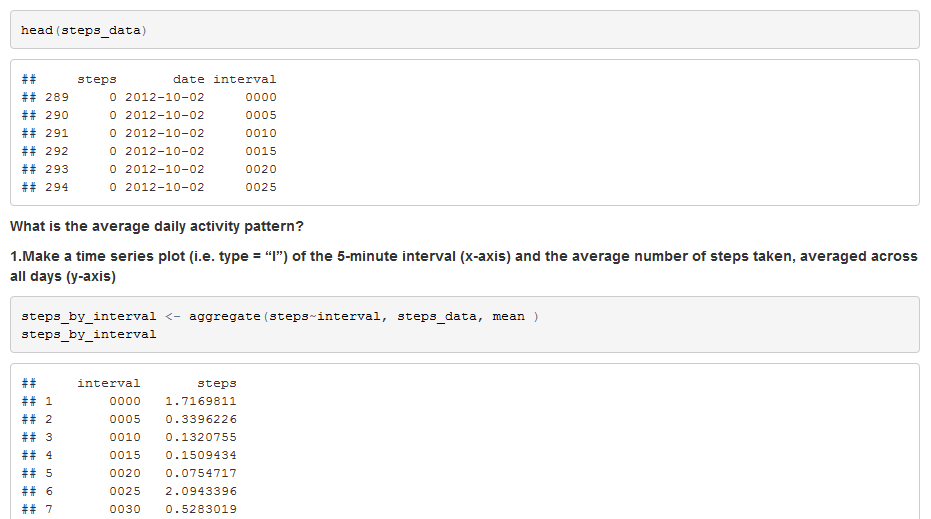
year Emissions

1 1999 7332967

2 2002 5635780

3 2005 5454703

4 2008 3464206



### Aggregate multiple columns

fatalities\_by\_event\_type = aggregate(cbind(FATALITIES, INJURIES)~EVTYPE, storm\_data, sum)

head(fatalities\_by\_event\_type)

## EVTYPE FATALITIES INJURIES

## 1 HAIL 4 225

## 2 TORNADO 2609 44405

## 3 TSTM WIND 148 2014

## Append a column to a data.frame

Note: The column data is inside a vector

|  |
| --- |
| > str(energy\_use\_subset)  'data.frame': 2880 obs. of 9 variables:  $ Date : chr "1/2/2007" "1/2/2007" "1/2/2007" "1/2/2007" ...  $ Time : chr "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...  $ Global\_active\_power : num 0.326 0.326 0.324 0.324 0.322 0.32 0.32 0.32 0.32 0.236 ...  $ Global\_reactive\_power: num 0.128 0.13 0.132 0.134 0.13 0.126 0.126 0.126 0.128 0 ...  $ Voltage : num 243 243 244 244 243 ...  $ Global\_intensity : num 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1 ...  $ Sub\_metering\_1 : num 0 0 0 0 0 0 0 0 0 0 ...  $ Sub\_metering\_2 : num 0 0 0 0 0 0 0 0 0 0 ...  $ Sub\_metering\_3 : num 0 0 0 0 0 0 0 0 0 0 ...  > str(dt)  POSIXlt[1:2880], format: "2007-02-01 00:00:00" "2007-02-01 00:01:00" "2007-02-01 00:02:00" ...  > energy\_use\_subset$date\_time <- dt  > str(energy\_use\_subset)  'data.frame': 2880 obs. of 10 variables:  $ Date : chr "1/2/2007" "1/2/2007" "1/2/2007" "1/2/2007" ...  $ Time : chr "00:00:00" "00:01:00" "00:02:00" "00:03:00" ...  $ Global\_active\_power : num 0.326 0.326 0.324 0.324 0.322 0.32 0.32 0.32 0.32 0.236 ...  $ Global\_reactive\_power: num 0.128 0.13 0.132 0.134 0.13 0.126 0.126 0.126 0.128 0 ...  $ Voltage : num 243 243 244 244 243 ...  $ Global\_intensity : num 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1 ...  $ Sub\_metering\_1 : num 0 0 0 0 0 0 0 0 0 0 ...  $ Sub\_metering\_2 : num 0 0 0 0 0 0 0 0 0 0 ...  $ Sub\_metering\_3 : num 0 0 0 0 0 0 0 0 0 0 ...  $ date\_time : POSIXlt, format: "2007-02-01 00:00:00" "2007-02-01 00:01:00". |
|  |
|  |

## Append a row to a data.frame

> df

id nobs

1 1 2

> df <- rbind(df, data.frame(id=17, nobs=34))

> df

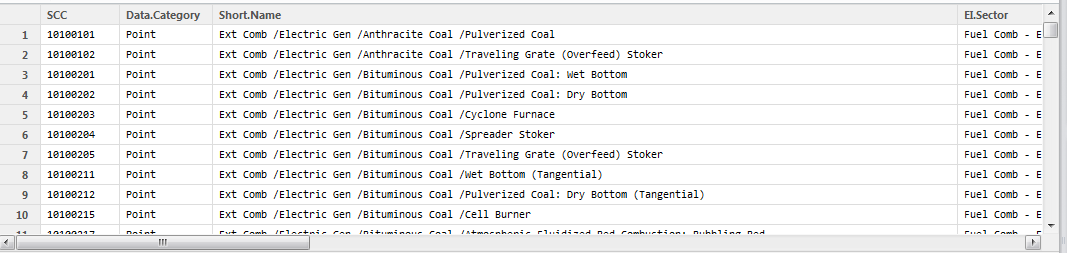
id nobs

1 1 2

2 17 34

## Browse a data.frame

View(SCCT)



## Convert a column to date data type

> energy\_use$Date <- as.Date(energy\_use$Date, format ="%d/%m/%Y")

> str(energy\_use)

'data.frame': 5258 obs. of 9 variables:

$ Date : Date, format: "2007-02-01" "2007-02-01" "2007-02-01" ...

## Create a data.frame from scratch

> df <- data.frame(id=integer(), nobs=integer())

> df <- rbind(df, data.frame(id=1, nobs=2))

> df

id nobs

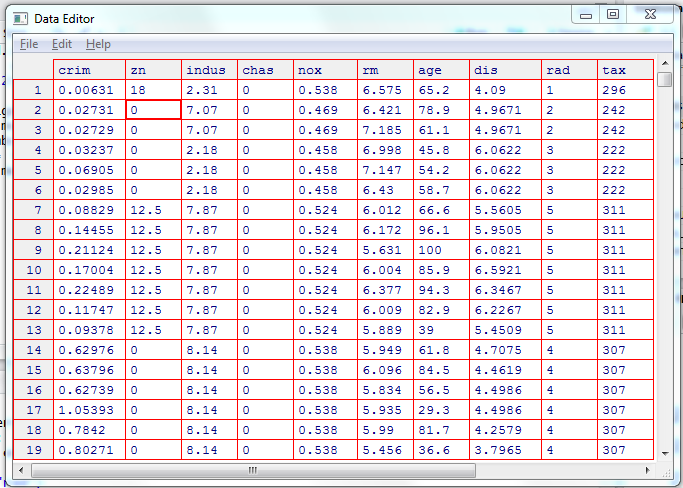
1 1 2

## Create a factor column based upon values in an existing column



## Edit a data.frame in table form

> fix(Boston)



## Expose the fields in a data.frame to the development environment

attach(Boston)

## Get a cell of data from a data.frame

> pollutantData[1,]

Date sulfate nitrate ID

1 2003-01-01 NA NA 1

> pollutantData[1,"ID"]

[1] 1

## Get the column names of a data.frame

> colnames(outcomeData)

## Get the number of rows in a data.frame

|  |
| --- |
| > nrow(data1)  [1] 153 |
|  |
| |  | | --- | |  | |

## Get a Row of Data from a data.frame

|  |
| --- |
| data1[1,]  Ozone Solar.R Wind Temp Month Day  1 41 190 7.4 67 5 1 |
|  |
| |  | | --- | |  | |

## Import a data.frame column as a formatted date

setClass('myDate')

setAs("character","myDate", function(from) as.Date(from, format="%d/%m/%Y") )

energy\_use <- read.csv("data/household\_power\_consumption\_short.txt",

sep=";", colClasses= c("myDate", "character", "numeric",

"numeric","numeric","numeric",

"numeric","numeric","numeric"))

## Look at the first part of a dataframe

head(df[whichrows, whichcols])

## Look at the structure of a data.frame

> str(energy\_use)

'data.frame': 5258 obs. of 9 variables:

$ Date : Factor w/ 4 levels "1/2/2007","2/2/2007",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Time : Factor w/ 1440 levels "00:00:00","00:01:00",..: 1 2 3 4 5 6 7 8 9 10 ...

$ Global\_active\_power : num 0.326 0.326 0.324 0.324 0.322 0.32 0.32 0.32 0.32 0.236 ...

$ Global\_reactive\_power: num 0.128 0.13 0.132 0.134 0.13 0.126 0.126 0.126 0.128 0 ...

$ Voltage : num 243 243 244 244 243 ...

$ Global\_intensity : num 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1.4 1 ...

$ Sub\_metering\_1 : num 0 0 0 0 0 0 0 0 0 0 ...

$ Sub\_metering\_2 : num 0 0 0 0 0 0 0 0 0 0 ...

$ Sub\_metering\_3 : num 0 0 0 0 0 0 0 0 0 0 ...

## Remove rows containing NA values from a data.frame

steps\_data <- na.omit(steps\_data)

## Select Rows which meet criteria

> test <- iris[iris$Species=="virginica", ]

> test

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

101 6.3 3.3 6.0 2.5 virginica

102 5.8 2.7 5.1 1.9 virginica

103 7.1 3.0 5.9 2.1 virginica

Data2<-Data[which(Data$Date %in% c("01/02/2007","02/02/2007")),]

## Select Rows which meet criteria using sqldf

setwd("D:/One Drive/Coursera/Exploratory Data Analysis/Assignments/Week 1/Data/exdata-data-household\_power\_consumption")

install.packages("sqldf")

library(sqldf)

## Read subset of data using read.csv.sql

data<-read.csv.sql("household\_power\_consumption.txt",header=TRUE,sep=";",

sql="Select \* from file where Date='1/2/2007' OR Date='2/2/2007'")

## Select Columns from a data.frame by index

outcomeData <- outcomeData[, c(2, 7, column)]

Here, columns 2, 7, and one additional column are retained

## Subsetting rows selecting on more than one value

We use the %in% notation when we want to subset on multiple values of y. The x.sub5 data frame contains only the observations for which the values of variable y are equal to either 1 or 4.

x.sub5 <- x.df[x.df$y %in% c(1, 4), ]

x.sub5

V1 V2 V3 V4 V5 y

1 -1.6862356 1.395021 1.35898920 1.8492410 1.7536886 1

2 0.8610318 -0.569828 -0.01984841 0.3570547 -0.9326248 1

5 0.6296957 1.794336 2.16226397 0.1604166 0.3721850 4

# Data Types

## Get the type of an object

|  |
| --- |
| > x <- 4  > class(x)  [1] "numeric" |
|  |
|  |

# Dates and Times

## Convert a String to a Date

> strptime("1/2/2007 00:01:00", "%d/%m/%Y %H:%M:%S")

[1] "2007-02-01 00:01:00 PST"

> class(strptime("1/2/2007 00:01:00", "%d/%m/%Y %H:%M:%S"))

[1] "POSIXlt" "POSIXt"

# Distributions

## Beta

### Cumulative Probability for a Beta Distribution

> pbeta(0.5, 5529, 5471, lower.tail = TRUE)

[1] 0.2901213

Note: This is the posterior probability that a die is fair (29% chance)

# Environment

## Change the Working Directory

> setwd("c:/r/Prog3")

> getwd()

[1] "c:/r/Prog3"

## Get the Working Directory

> getwd()

[1] "C:/Users/Bruce/Documents"

>

## List the Files in the Working Directory

dir()

## Load a Function into the Environment

source(“myCode.R”)

## List the Functions Available in the Environment

ls()

# Files and Directories

## Create a directory

> dir.create("data")

## Download a file from the web

download.file()

## See if a directory or file exists

> file.exists("no such dir")

[1] FALSE

# Libraries

## Loading Libraries

> library(ISLR)

# Linear Regression

## Confidence interval for a regression line fit

> confint(fit)

2.5 % 97.5 %

(Intercept) 33.448457 35.6592247

lstat -1.026148 -0.8739505

## Generate the Regression line

> fit = lm(medv~lstat, Boston)

> fit

Coefficients:

(Intercept) lstat

34.55 -0.95

Note generates the regression line for mdev as a function of lstat

## Get more information about the linear regression

> summary(fit)

Residuals:

Min 1Q Median 3Q Max

-15.168 -3.990 -1.318 2.034 24.500

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 34.55384 0.56263 61.41 <2e-16 \*\*\*

lstat -0.95005 0.03873 -24.53 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

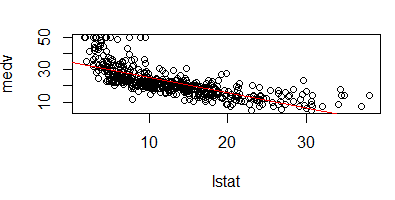
Residual standard error: 6.216 on 504 degrees of freedom

Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432

F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

## Plot a regression line on a plot

> abline(fit, col="red")



# Matrices

## Create a matrix

x <- matrix(c(1,2,3,4), 2,2)

> x

[,1] [,2]

[1,] 1 3

[2,] 2 4

# Misc

## Printing a variable to the console

for (i in id) {

**cat(i, "\n")**

}

# Packages

## Installing packages

> install.packages("ISLR")

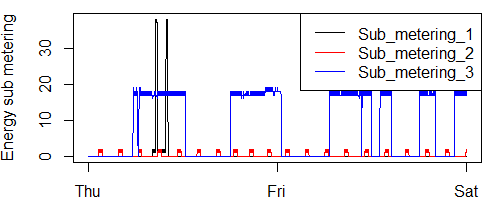
# Plotting data

## Adding a legend to a base plot

legend(x='topright',

legend=c("Sub\_metering\_1 ", "Sub\_metering\_2 ", "Sub\_metering\_3 "),

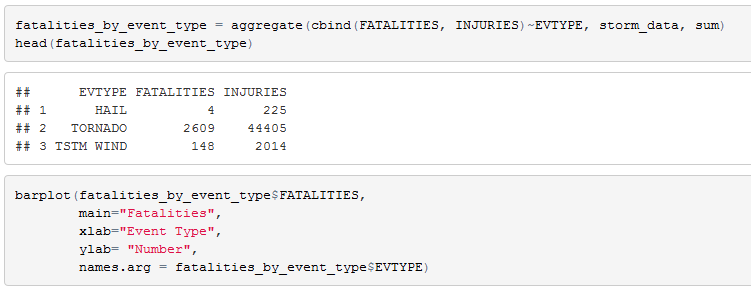
col=c('black', 'red', 'blue'), lty=c(1,1,1), cex=0.5 )

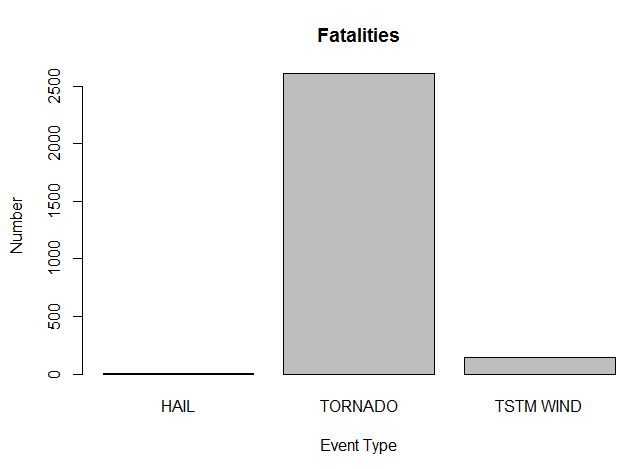


lty = line type. 0=blank, 1=solid

cex : Sets relative size of the legend

## Bar Plots





## faceted plots (ggplot2)

str(balt)

'data.frame': 2096 obs. of 6 variables:

$ fips : chr "24510" "24510" "24510" "24510" ...

$ SCC : chr "10100601" "10200601" "10200602" "30100699" ...

$ Pollutant: chr "PM25-PRI" "PM25-PRI" "PM25-PRI" "PM25-PRI" ...

$ Emissions: num 6.53 78.88 0.92 10.38 10.86 ...

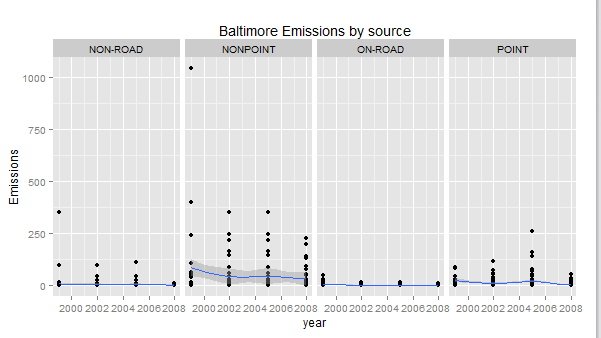
$ type : chr "POINT" "POINT" "POINT" "POINT" ...

$ year : int 1999 1999 1999 1999 1999 1999 1999 1999 1999 1999 ...

qplot(year, Emissions, data=balt, facets=.~type,

geom=c("point", "smooth"), method="loess") +

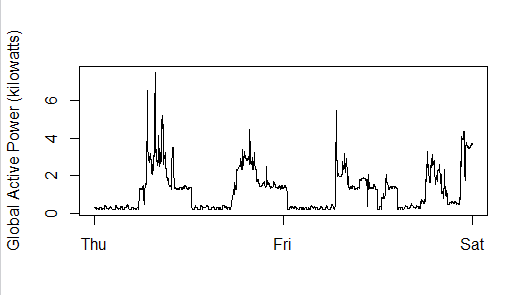
labs(title="Baltimore Emissions by source")



## Line Plots

plot(energy\_use\_subset$date\_time, energy\_use\_subset$Global\_active\_power, type="l",

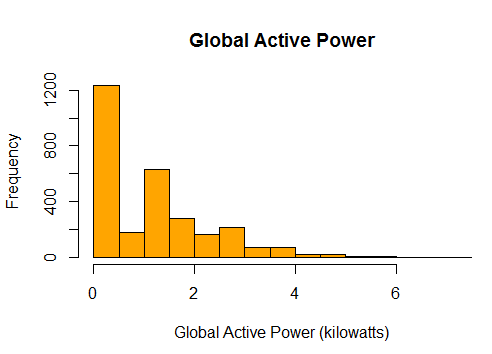
xlab='', ylab='Global Active Power (kilowatts)')



## Histograms

hist(energy\_use\_subset$Global\_active\_power, main='Global Active Power', col='orange',

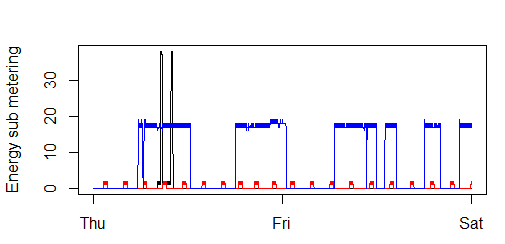
xlab='Global Active Power (kilowatts)')



## Line plots with multiple lines

plot(x,y1,type="l",col="red")

lines(x,y2,col="green")



## Multiple plots in a grid

require(gridExtra)

plot1 <- qplot(1)

plot2 <- qplot(1)

grid.arrange(plot1, plot2, ncol=2)

## Multiple plots in quadrants

par(mfrow=c(2,2), mar=c(4,4,2,1), oma=c(0,0,2,0))

with(energy\_use\_subset, {

plot(energy\_use\_subset$date\_time, energy\_use\_subset$Global\_active\_power, type="l",

xlab='', ylab='Global Active Power (kilowatts)')

plot(energy\_use\_subset$date\_time, energy\_use\_subset$Voltage, type="l",

xlab='datetime', ylab='Voltage')

plot(energy\_use\_subset$date\_time, energy\_use\_subset$Sub\_metering\_1, type="l",

xlab='', ylab='Energy sub metering')

lines(energy\_use\_subset$date\_time, energy\_use\_subset$Sub\_metering\_2, col='red')

lines(energy\_use\_subset$date\_time, energy\_use\_subset$Sub\_metering\_3, col='blue')

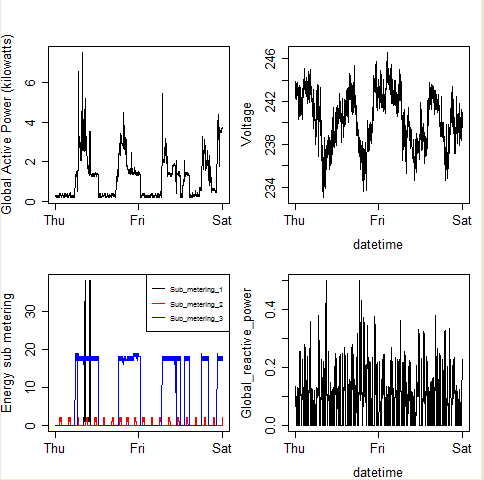
legend(x='topright', legend=c("Sub\_metering\_1 ", "Sub\_metering\_2 ", "Sub\_metering\_3 "),

col=c('black', 'red', 'blue'), lty=c(1,1,1), cex=0.5 )

plot(energy\_use\_subset$date\_time, energy\_use\_subset$Global\_reactive\_power, type="l",

xlab='datetime', ylab='Global\_reactive\_power')

})



## Plotting to a PNG file

hist(energy\_use\_subset$Global\_active\_power, main='Global Active Power', col='orange',

xlab='Global Active Power (kilowatts)')

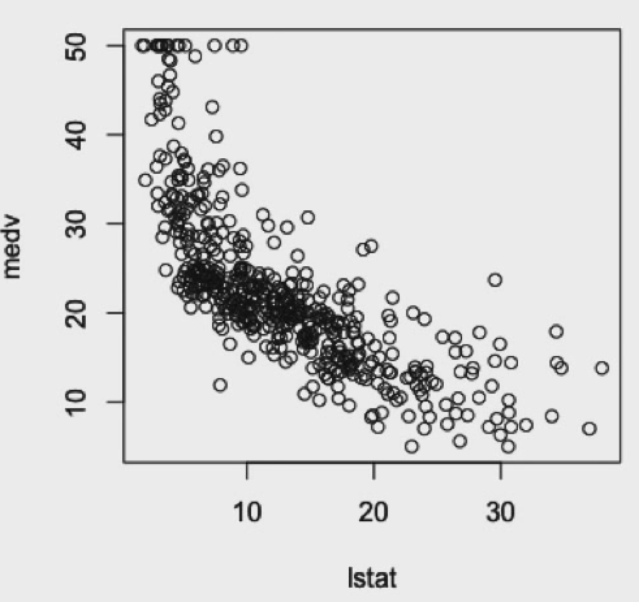
# Now copy the histogram to a png file

dev.copy(png, file='plot1.png', width = 480, height = 480 )

dev.off()

## Scatter plot of y vs x

plot(medv~lstat, Boston)



## Using Base Plot

NEI <- readRDS("summarySCC\_PM25.rds")

balt <- NEI[NEI$fips=="24510",]

yearlyEmissions <- aggregate(Emissions~year, balt, sum)

with(yearlyEmissions, plot(year, Emissions))

title(main= "Yearly Total PM2.5 Emissions for Baltimore")

# Reading Data

## Read a CSV file into a data.frame

> data1 <- read.csv("hw1\_data.csv")

> data1

Ozone Solar.R Wind Temp Month Day

1 41 190 7.4 67 5 1

2 36 118 8.0 72 5 2

3 12 149 12.6 74 5 3

## Read from an excel spreadsheet

library(xlsx)

read.xlsx()

# Statistics

## p-value for a Binomial Distribution

p(0, 1, 2, 3) successes in 20 trials, p=0.5

sum(dbinom(0:4, size=20, p=0.5))

# Strings

## Concatenate two strings

|  |
| --- |
| > paste("a","b", sep = "")  [1] "ab" |
|  |
| |  | | --- | | > |   > paste0("1","2")  [1] "12" |

## Padding a string with zeros

> sprintf("%03d", 17)

[1] "017"

steps\_data$interval = str\_pad(steps\_data$interval, 4, pad="0")

steps\_data

## steps date interval

## 289 0 2012-10-02 0000

## 290 0 2012-10-02 0005

## 291 0 2012-10-02 0010

## 292 0 2012-10-02 0015

# Vectors

## Concatenate two string vectors

> date\_part <- energy\_use\_subset[, 'Date']

> time\_part <- energy\_use\_subset[, 'Time']

> date\_time\_str <- paste(date\_part, time\_part, sep=' ')

> str(date\_time\_str)

chr [1:2880] "1/2/2007 00:00:00" "1/2/2007 00:01:00" "1/2/2007 00:02:00" ...

## Creating a vector

> a = 1:2

> a

[1] 1 2

|  |
| --- |
| > x <- c(1,3,2,5)  > x  [1] 1 3 2 5 |
|  |
| |  | | --- | |  | |

## Iterating over a vector

for (i in id) {

writeLines(paste(i))

}

|  |
| --- |
| 1  2 |
|  |
| |  | | --- | |  | |

## Length of a vector

> bad <- ozone\_col[is.na(ozone\_col)]

> length(bad)

[1] 37

## Return valid elements of a vector

|  |
| --- |
| > a  [1] 1 2 NA 4  > a[complete.cases(a)]  [1] 1 2 4 |
|  |
| |  | | --- | |  | |