DA-习题课

20210419

1. DataCamp
2. 课堂测试讲解

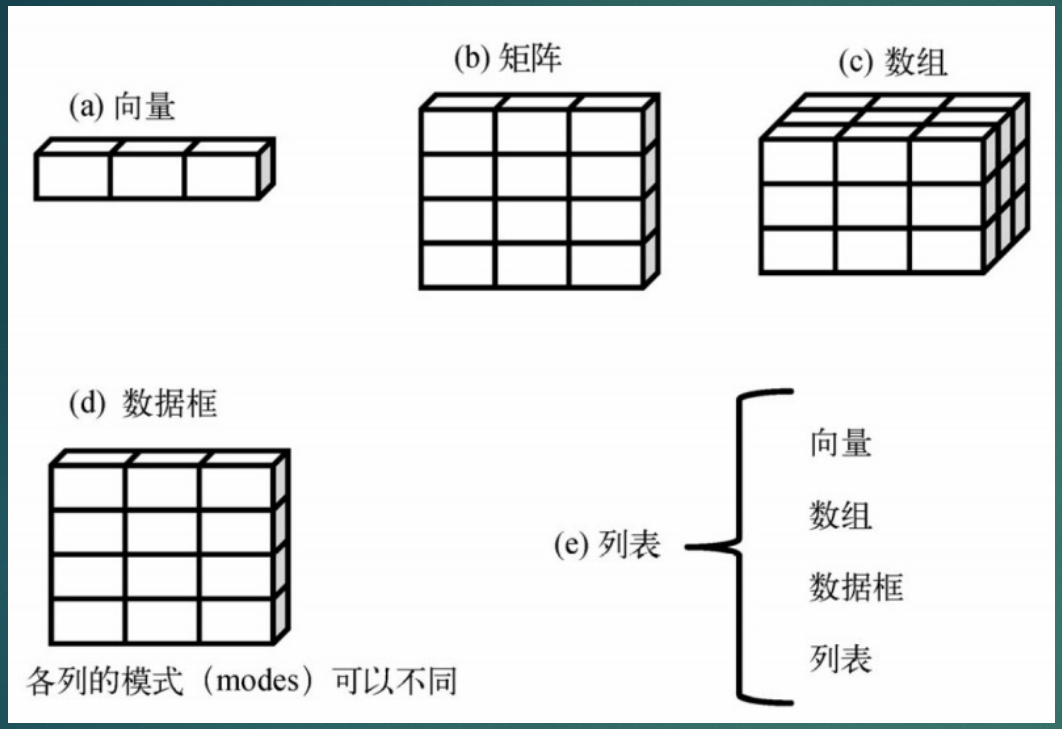
1.1 Introduction to R & Intermediate R

1. 基本数据结构的创建和使用
2. 基本运算
3. 常见函数的定义和使用
4. 自定义函数的使用与Apply族函数
5. 其他

更多详细内容参见课件DA-02和DA03

1.1 Introduction to R & Intermediate R

1. 基本数据结构的创建和使用



1.1 Introduction to R & Intermediate R

1. 基本数据结构的创建和使用

向量：

创建：v <- c(1,2,3); seq(from=value1, to=value2,by=value3); rep(x,times=n)

提取：x[5]; x[1:3], x[c(1,3)], x[-5]（去除）

矩阵：

创建：m <- matrix(vector, nrow, ncol, byrow, dimnames=list(rownames, colnames)

提取：m[i, j], m[i,], m[, j]

由矩阵可以引申到更高维的array

数据框：

不同的列可以包含不同模式的数据

创建：df <- data.frame(col1_vec, col2_vec, ...), coli_vec为列向量，可以是任意类型

提取：df[1:2], df$col1, df[c(“col1”,“col2”)]

列表:

一些数据对象的有序集合

创建：l <- list(object1, object2, ...); l <- list(name1=object1, name2=object2, ...)

提取：l[[1]], l$name

其他：

因子的创建和使用

str()查看数据对象的结构;

summary()查看数据对象的统计概要

1.1 Introduction to R & Intermediate R

2. 基本运算

+	加
−	减
*	乘
/	除
^, **	求幂
x %% Y	求余
x %/% Y	整除

>,<	大于，小于
<=, >=	大于等于，小于等于
!==, ==	不等于，等于
!x	非x
x | y	x或者y
x & y	x和y
isTRUE(x)	x是否为TRUE

3. 常见函数的定义和使用

详见DA-02 P16-P18 & DA-03 P05-P15部分

1.1 Introduction to R & Intermediate R

4. 自定义函数的使用与Apply函数族

```
function(arg1,arg2,...) {
        statements
        return (object)
        }
```

```
myfun <- function() {
                    print("hello world")
                    return ()
                   }
```

```
> f <- function(x,y) x + y
> f
function(x,y) x + y
> f(1,2)
[1] 3
```

1.1 Introduction to R & Intermediate R

4. 自定义函数的使用与Apply族函数

Description

Returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.

Usage

```
apply(X, MARGIN, FUN, …)
```

Arguments

`X`	an array, including a matrix.
`MARGIN`	a vector giving the subscripts which the function will be applied over. E.g., for a matrix `` `1` `` indicates rows, `` `2` `` indicates columns, `` `c(1, 2)` `` indicates rows and columns. Where `` `X` `` has named dimnames, it can be a character vector selecting dimension names.
`FUN`	the function to be applied: see 'Details'. In the case of functions like `` `+` ``, `` `%*%` ``, etc., the function name must be backquoted or quoted.
…	optional arguments to `` `FUN` ``.

1.1 Introduction to R & Intermediate R

4. 自定义函数的使用与Apply族函数

Apply()函数族除了Apply()函数外，常见的还包括lapply(), sapply()等函数。

lapply(X, FUN, ...)，lapply()函数返回一个与输入X等长度的list；
sapply(X, FUN, ...)，sapply是一个简化且使用更加友好的版本，其返回值是一个向量或矩阵

使用apply族函数的最大优势可以代替循环结构，简化编程并提高执行效率
更多详细的使用可在https://www.rdocumentation.org/中查询。

5. 其他

流程控制、循环语句等，参见DA-04。

1.2 Introduction to Importing Data in R

1. CSV文件
2. TXT文件
3. XLS文件
4. XLConnect

更多详细内容可参见https://www.tutorialspoint.com/r/

1.2 Introduction to Importing Data in R

R Language read.table()

表2-2 函数read.table()的选项

选 项	描 述
header	一个表示文件是否在第一行包含了变量名的逻辑型变量
sep	分开数据值的分隔符。默认是 sep=""，这表示了一个或多个空格、制表符、换行或回车。使用 sep=","来读取用逗号来分隔行内数据的文件，使用 sep="\t"来读取使用制表符来分割行内数据的文件
row.names	一个用于指定一个或多个行标记符的可选参数
col.names	如果数据文件的第一行不包括变量名（header=FASLE），你可以用 col.names 去指定一个包含变量名的字符向量。如果 header=FALSE 以及 col.names 选项被省略了，变量会被分别命名为 V1、V2，以此类推
na.strings	可选的用于表示缺失值的字符向量。比如说，na.strings=c("-9", "?")把-9 和?值在读取数据的时候转换成 NA
colClasses	可选的分配到每一列的类向量。比如说，colClasses=c("numeric", "numeric", "character", "NULL", "numeric")把前两列读取为数值型变量，把第三列读取为字符型向量，跳过第四列，把第五列读取为数值型向量。如果数据有多余五列，colClasses 的值会被循环。当你在读取大型文本文件的时候，加上 colClasses 选项可以可观地提升处理的速度
quote	用于对有特殊字符的字符串划定界限的自负床。默认值是双引号（"）或单引号（'）
skip	读取数据前跳过的行的数目。这个选项在跳过头注释的时候比较有用
stringsAsFactors	一个逻辑变量，标记处字符向量是否需要转化成因子。默认值是 TRUE，除非它被 colClases 所覆盖。当你在处理大型文本文件的时候，设置成 stringsAsFactors=FALSE 可以提升处理速度
text	一个指定文字进行处理的字符串。如果 text 被设置了，file 应该被留空。2.3.1 节给出了一个例子

1.2 Introduction to Importing Data in R

1. CSV文件

“逗号分隔值（Comma-Separated Values，CSV，有时也称为字符分隔值，因为分隔字符也可以不是逗号），其文件以纯文本形式存储表格数据（数字和文本）”——百度百科

read.csv()

readr: read_csv(file, skip, nmax)

data.table: fread(file, select, drop)

2. TXT文件

read.delim() 注意header参数和colNames参数的搭配

readr: read_delim(file, delim, quote = "\", col_names = TRUE, na = c("", "NA"), skip = 0, n_max = Inf)

1.2 Introduction to Importing Data in R

3. XLS文件

readxl库：

excel_sheets(filePath)

获取xls文件中的每个sheet名

read_excel(filePath, sheet=sheetName or sheetNumber)

读取xls文件中具体某个sheet信息

Tips: 可以利用lapply函数读取所有sheet提高执行效率

read_excel常用的可选参数：

col_names：FALSE or vector，是否指定列名

skip: 读取数据前跳过若干行

gdata库:

read.xls(path, sheet=sheetName or sheetNumber)

1.2 Introduction to Importing Data in R

4. XLConnect库的使用

XLConnect库封装一系列接口来实现对XLS文件的读写，主要包括一些方法：

```
myBook <- loadWorkbook(filepath)
getSheets(myBook)
readWorksheet(my_book, sheet = Name/Number, startCol, endCol)

createSheet(my_book, sheetName)
writeWorksheet(my_book, sheet, sheetName)
saveWorkbook(my_book, file =  “xxx.xlsx")

renameSheet(my_book, "data_summary", "summary")
removeSheet(my_book, 4)
```

1.3 Cleaning Data in R

1. 数据类型转换
2. 重复值处理
3. 异常值处理

1.3 Cleaning Data in R

1. 数据类型转换

glimpse() 查看数据框中所有列的数据类型

summary() 查看数据框概括信息

利用mutate()向数据框增加一列的数据：

```
df <- mutate(df, newscore = score2020)
OR   df <- df %>% mutate(newscore = score2020)
```

(%>%，管道，提高编码的效率和可读性）

利用as.TYPE()进行数据类型转换

1.3 Cleaning Data in R

2. 重复值处理

duplicated()

distinct()

df %>% count(col) %>% filter(n > 1)

3. 异常值/缺失值处理

根据实际数据来判断是否异常，eg：年月日、时间

根据数据整体分布判断是否异常， eg：绘图查看

is.na()

filter(!is.na(vec))

1.4 Data Visualization & Visualizing Time Series Data in R

此部分为近期课程学习内容，如有疑问请参见课件DA-05 ~ DA-10

基础: plot(x, y, ...)

`main`

an overall title for the plot: see `title`.

`sub`

a sub title for the plot: see `title`.

`xlab`

a title for the x axis: see `title`.

`ylab`

a title for the y axis: see `title`.

`type`

what type of plot should be drawn. Possible types are

- `"p"` for points,
- `"l"` for lines,
- `"b"` for both,
- `"c"` for the lines part alone of `"b"`,
- `"o"` for both 'overplotted',
- `"h"` for 'histogram' like (or 'high-density') vertical lines,
- `"s"` for stair steps,
- `"S"` for other steps, see 'Details' below,
- `"n"` for no plotting.

1.4 Data Visualization & Visualizing Time Series Data in R

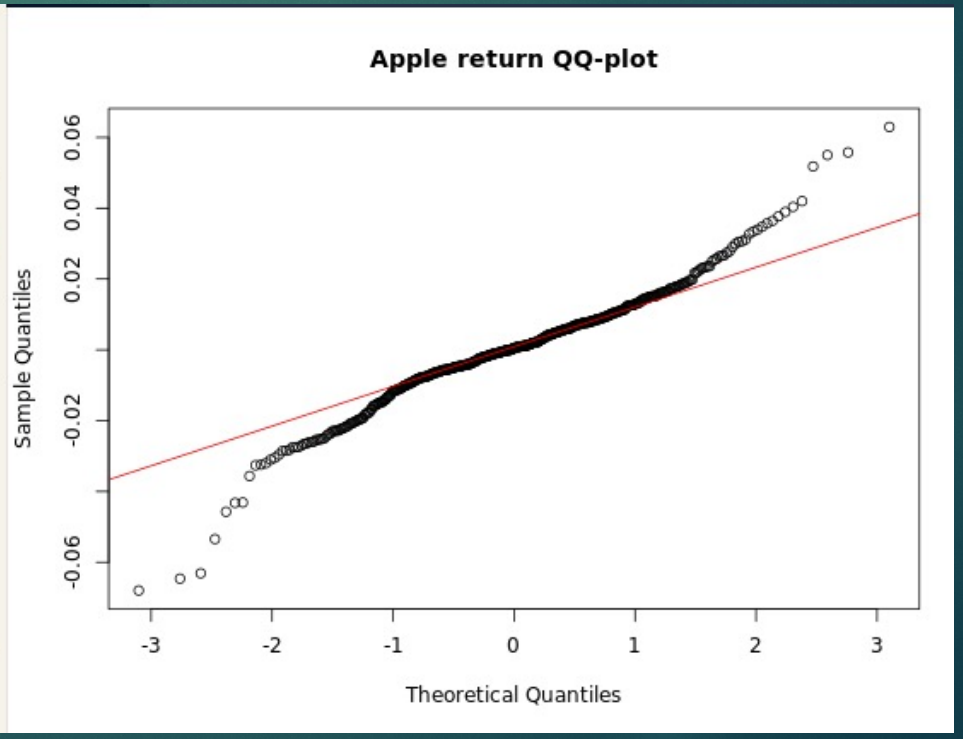
此部分为近期课程学习内容，如有疑问请参见课件DA-05 ~ DA-10

正态分布数据的展示：q-qplot

用qqnorm()和qqline()作正态QQ图。 当变量样本来自正态分布总体时，图的散点近似在一条直线周围。

```
# Create q-q plot
qqnorm(rtn, main = "Apple return QQ-plot")

# Add a red line showing normality
qqline(rtn, col = "red")
```



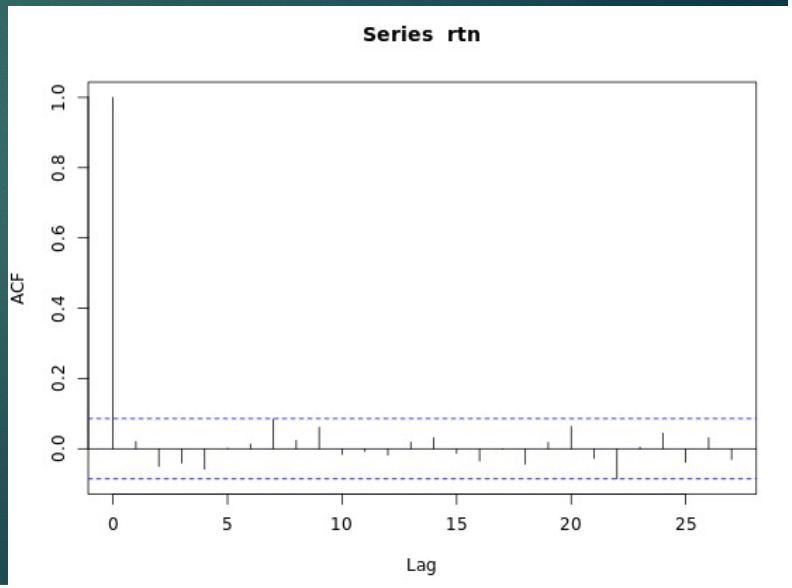
1.4 Data Visualization & Visualizing Time Series Data in R

此部分为近期课程学习内容，如有疑问请参见课件DA-05 ~ DA-10

时序数据相关性展示：

acf(x, lag.max = NULL, type = c("correlation", "covariance", "partial"),
plot = TRUE, na.action = na.fail, demean = TRUE, ...)

x: 一元或多元（不ccf）数字时间序列对象或一个数值向量或矩阵
lag.max: 计算acf的最大延迟， 默认值为10*log10(N/m)，其中N是观察数，m是序列维度
type: 指定具体的相关性计算指标，包括自相关、协方差和偏自相关
plot: 是否绘图
na.action: 缺失值处理方法



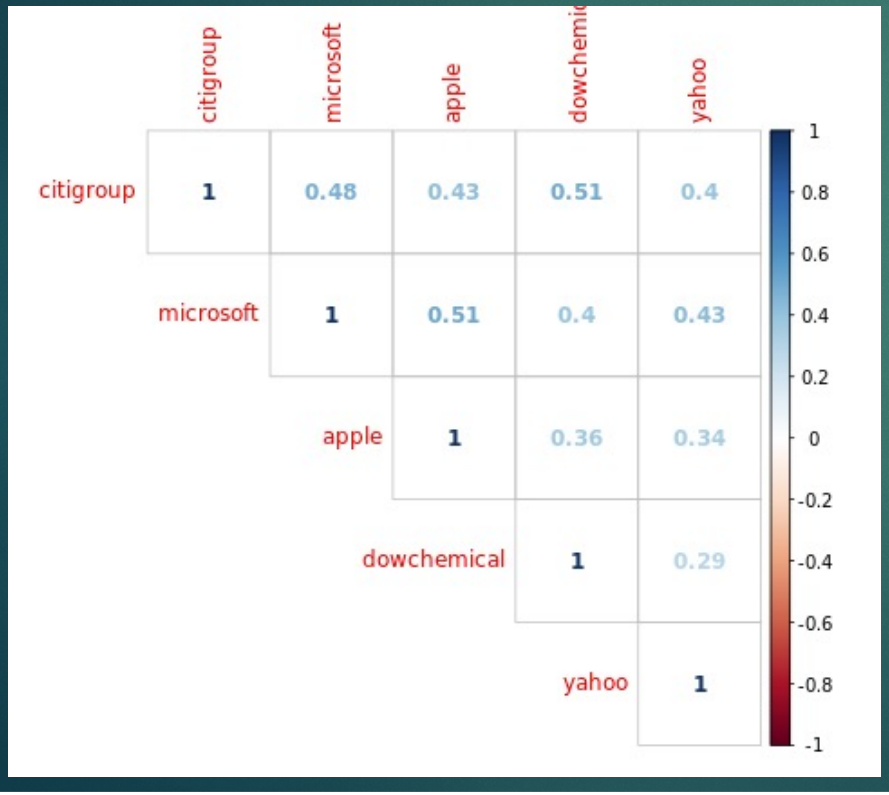
1.4 Data Visualization & Visualizing Time Series Data in R

此部分为近期课程学习内容，如有疑问请参见课件DA-05 ~ DA-10

相关性矩阵及其展示

cov(x, y = NULL, method = c("pearson", "kendall", "spearman"))

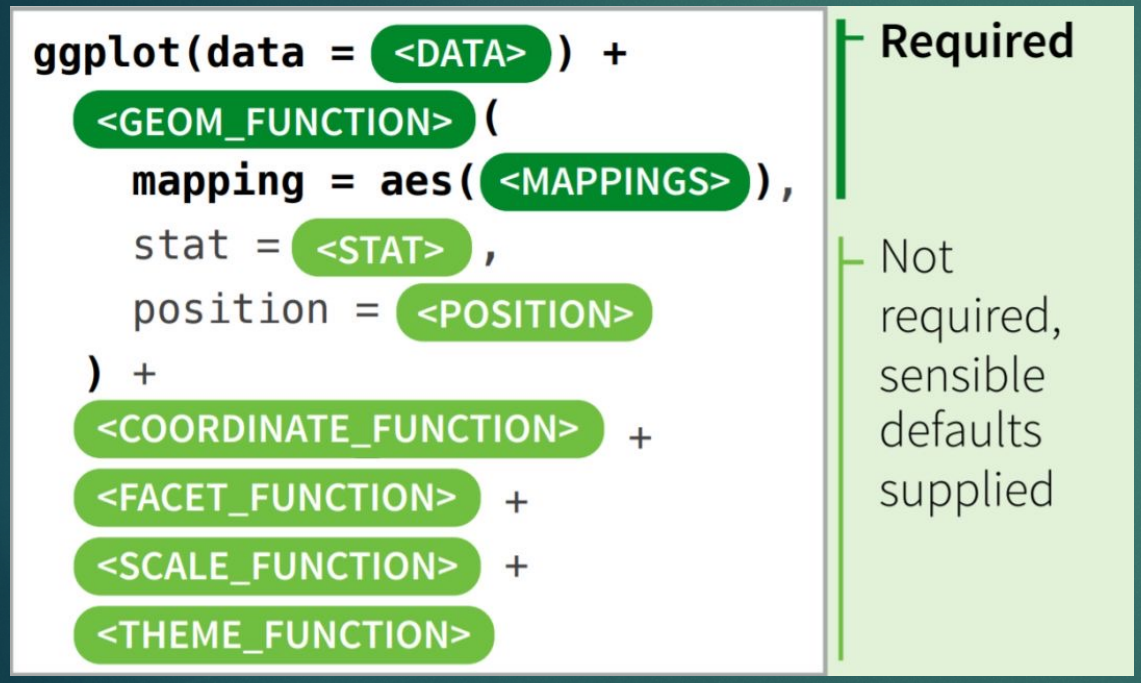
corplot(M, ...)



1.4 Data Visualization & Visualizing Time Series Data in R

此部分为近期课程学习内容，如有疑问请参见课件DA-05 ~ DA-10

ggplot2绘图的整体流程



2. 课堂测试讲解

R Graphics II 课堂测试04 *30分钟*

1. 使用鸢尾花数据iris

1）先用names()观察其结构，然后用花瓣长度和宽度做散点图

2）在plot函数里面添加细节。修改点的形状和颜色由白色空心圆换成红色雪花；修改坐标轴名称并添加标题 "relationship between width and length of Iris petal"。

2. 使用 airquality 数据

1）绘温度 Temp 直方图，加一个横坐标"Temperature",加一个标题"The Distribution of Temperature“

2）频数变频率，并设置颜色为绿色

3）四幅图放在一个面板里，两个一排。并使用MASS包的trueHist函数画出频率直方图：

第一幅图，airquality里温度变量的直方图（频数）

第二幅图，airquality里该变量的直方图（频率）并添加密度曲线，填充红色

第三幅图，airquality里风速变量的直方图（频数）

第四幅图，airquality里该变量的直方图（频率），并添加密度曲线，填充蓝色

3. 使用mtcars里的mpg做箱图

给箱图添加坐标轴：x轴为”Number of Cylinders”，y轴为=“Miles Per Gallon”标题”Car Milage Data”。根据不同cyl变量下mpg的箱线图，并添加x轴”Number of Cylinders”,y轴”Miles Per Gallon”

4. 按要求作图：

1) 创建字符向量colors,元素为"green","orange","brown“;创建字符向量months,元素为"一月","二月","三月","四月","五月“;创建字符向量regions,元素为"东部地区","西部地区","南部地区“;创建矩阵values,元素为值2,9,3,11,9,4,8,7,3,12,5,2,8,10,11，要求3行5列

2) 使用矩阵values创建推叠的条形图，添加标题为”总收入”，x轴名称为”月份”，y轴名称为”收入”，条形图的标签为字符向量months(使用names.arg参数)，推叠台型图的颜色设置为创建的字符向量colors

3) 添加图例，内容为字符向量regions，分别对应条形图中的三种颜色

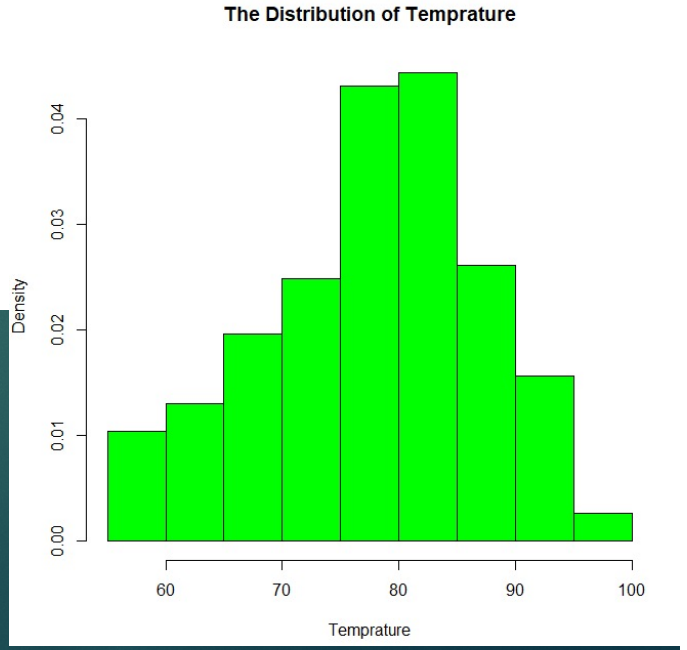
2.1 课堂测试4

1.

```
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
>
>
>
>
> plot(iris$Speal.Length, iris$Speal.Width, xlab="length", ylab="width", col="red", pch=8, main="relation ship between width and length of its petal")
```

2.

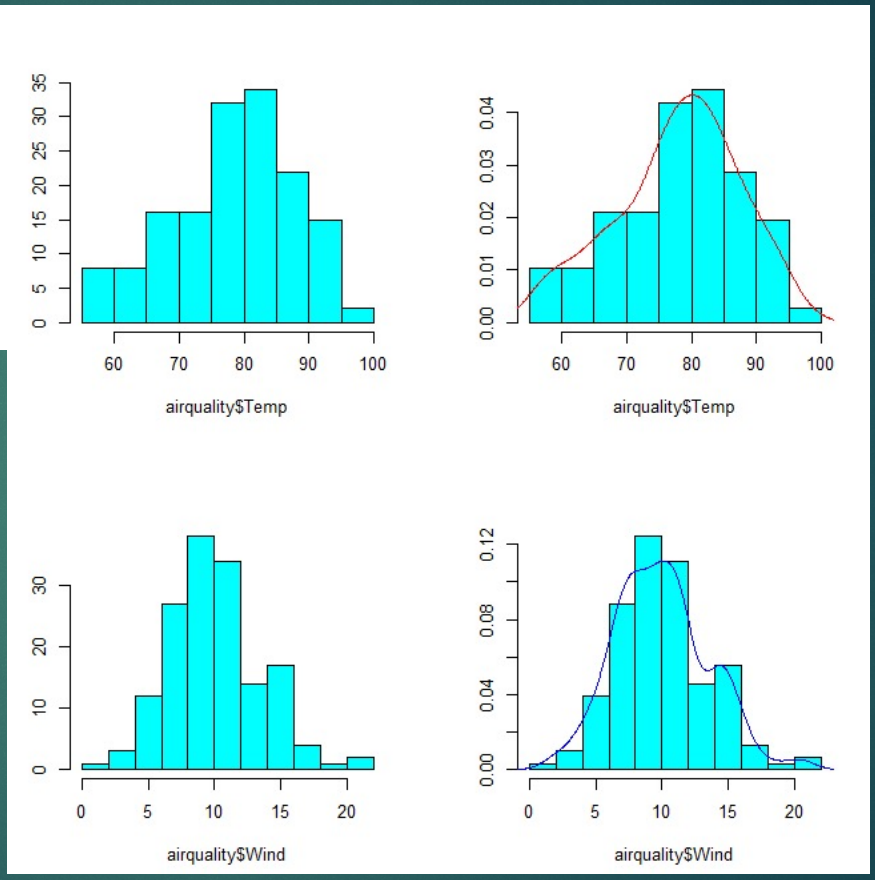
```
> names(airquality)
[1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
> str(airquality)
'data.frame':   153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
>
> hist(airquality$Temp, xlab="Temprature", main="The Distribution of Temprature", col="green", freq=F)
```



2.1 课堂测试4

2.

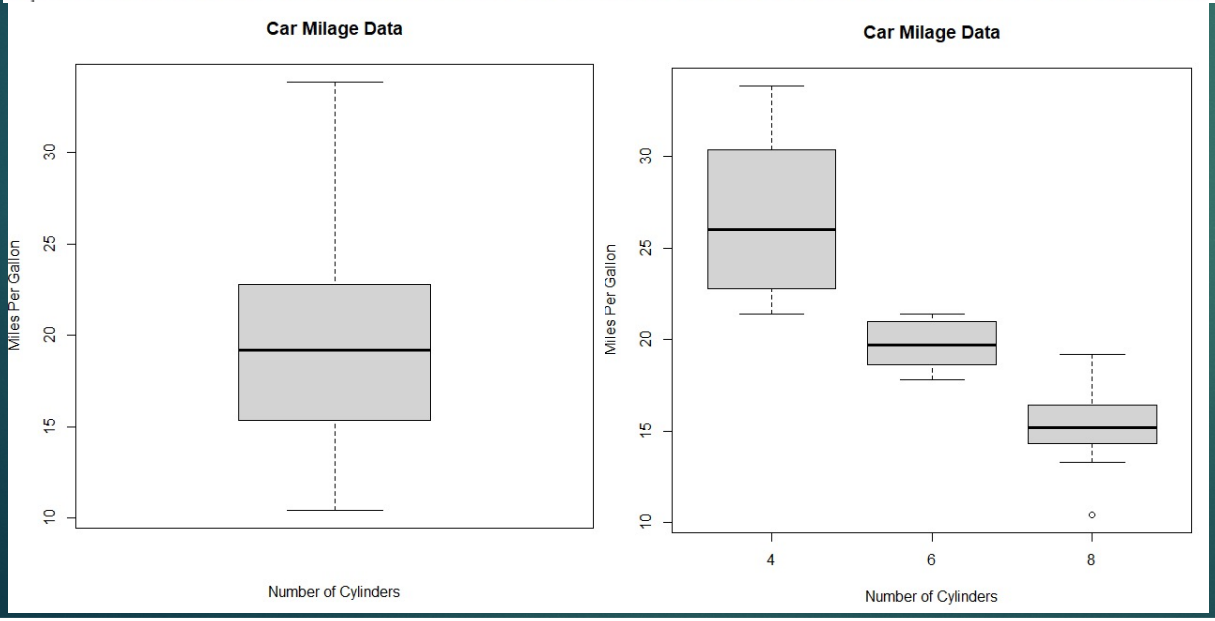
```
> library(MASS)
>
> par(mfrow=c(2,2))
> truehist(airquality$Temp, prob=FALSE)
> truehist(airquality$Temp, prob=TRUE)
> lines(density(airquality$Temp), col="red")
> truehist(airquality$Wind, prob=FALSE)
> truehist(airquality$Wind, prob=TRUE)
> lines(density(airquality$Wind), col="blue")
>
```



2. 课堂测试4

3.

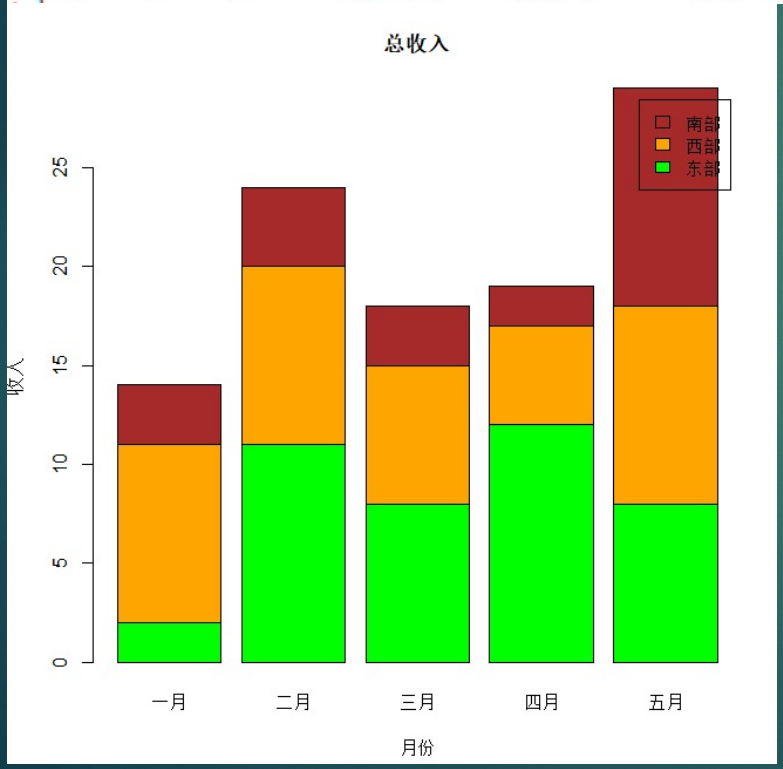
```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
>
>
> boxplot(mtcars$mpg, xlab="Number of Cylinders", ylab="Miles Per Gallon", main="Car Milage Data")
> boxplot(mtcars$mpg ~ mtcars$cyl, xlab="Number of Cylinders", ylab="Miles Per Gallon", main="Car Milage Data")
```



2. 课堂测试4

4.

```
> colors <- c("green", "orange", "brown")
> months <- c("一月", "二月", "三月", "四月", "五月")
> regions <- c("东部", "西部", "南部")
> values <- matrix(c(2, 9, 3, 11, 9, 4, 8, 7, 3, 12, 5, 2, 8, 10, 11), nrow=3)
>
>
> barplot(values, xlab="月份", ylab="收入", main="总收入", col=colors, names.arg=months, legend=regions)
```



2.2 课堂测试5

Course Wrap-up

课堂测试05

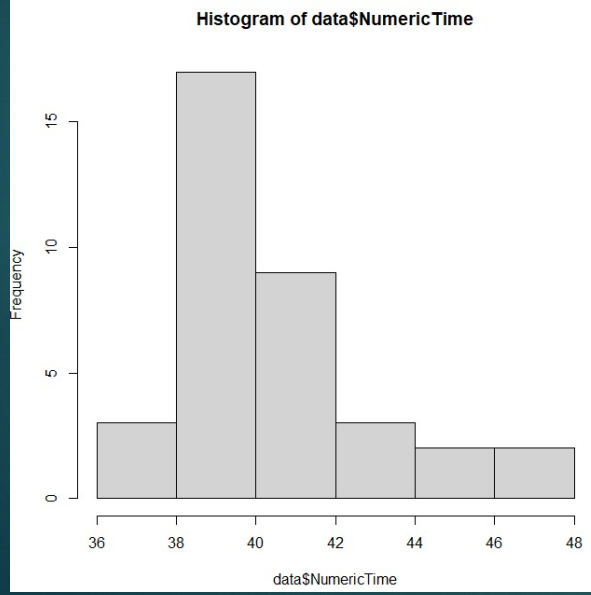
30分钟

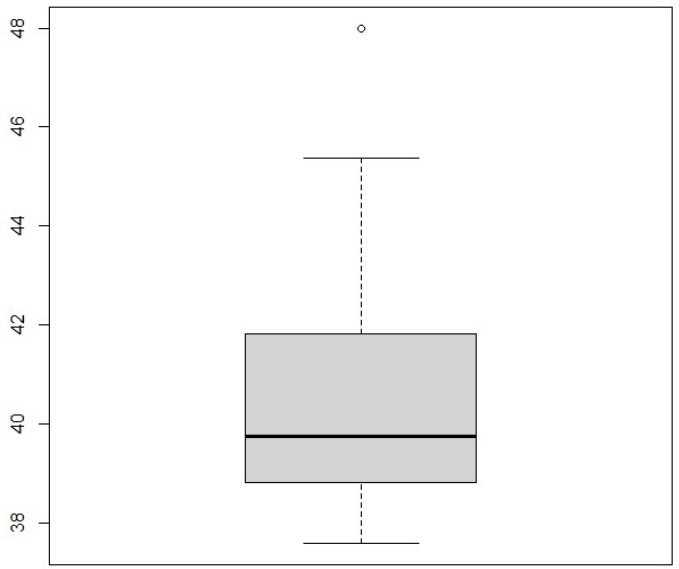
- 1、数据集alpe_d_huez2描述了环法自行车赛期间Alpe d'Huez赛段的最快时间，以及关于年份和吸毒指控的背景信息。绘制出车手最快时间的分布。使用a）直方图和b）箱线图显示它们。
- 2、mtcars是datasets包中的数据集。请使用str()函数了解这个数据集的构成，并输出数据集，然后按要求画图：
  - a. 我们要设置一个蓝色背景和红色的点或线。 我们应该使用什么命令
  - b. 画出cyl和mpg关系的散点图，并将结果输出为plot.png，要求输出为白底，360px*360px,点的大小为72
- 3、obama_vs_mccain数据集描述了2008年美国总统选举中的各州投票信息，以及关于收入，失业，种族和宗教的背景信息。
  - a. 画出收入Income和参加选举比例Turnout之间的关系的散点图。提示：Turnout存在Na值。
  - b. 将上述图形点的形状为黑色实心三角形(17)
  - c. 数据集中有一个因子类型的列regions,请画出每个地区region下的收入Income和参加选举比例Turnout之间的关系的散点图。要求设置布局为5列，行优先。

2.2 课堂测试5

1.

```
> getwd()
[1] "C:/Users/Melyn/Documents"
> data <- read.csv("alpe_d_huez2.csv")
> str(data)
'data.frame':   36 obs. of  7 variables:
 $ Time       : chr  "37' 35\"" "37' 36\"" "38' 00\"" "38' 01\"" ...
 $ NumericTime: num  37.6 37.6 38 38 38.1 ...
 $ Name       : chr  "Marco Pantani" "Lance Armstrong" "Marco Pantani" "Lance Armstrong" ...
 $ Year       : int  1997 2004 1994 2001 1995 1997 2006 2006 2004 1997 ...
 $ Nationality: chr  "Italy" "United States" "Italy" "United States" ...
 $ DrugUse    : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ Allegations: chr  "Alleged drug use during 1997 due to high haematocrit levels." "2004 Tour
> hist(data$NumericTime)
> boxplot(data$NumericTime)
```

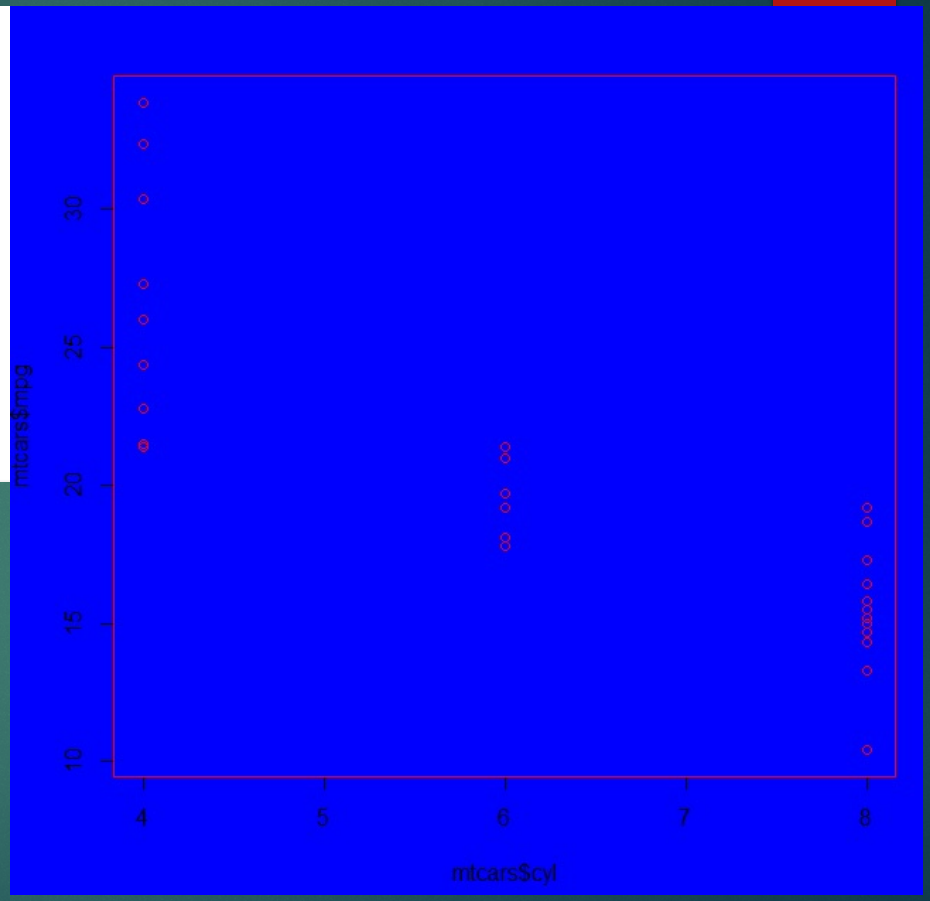




2.2 课堂测试5

2.

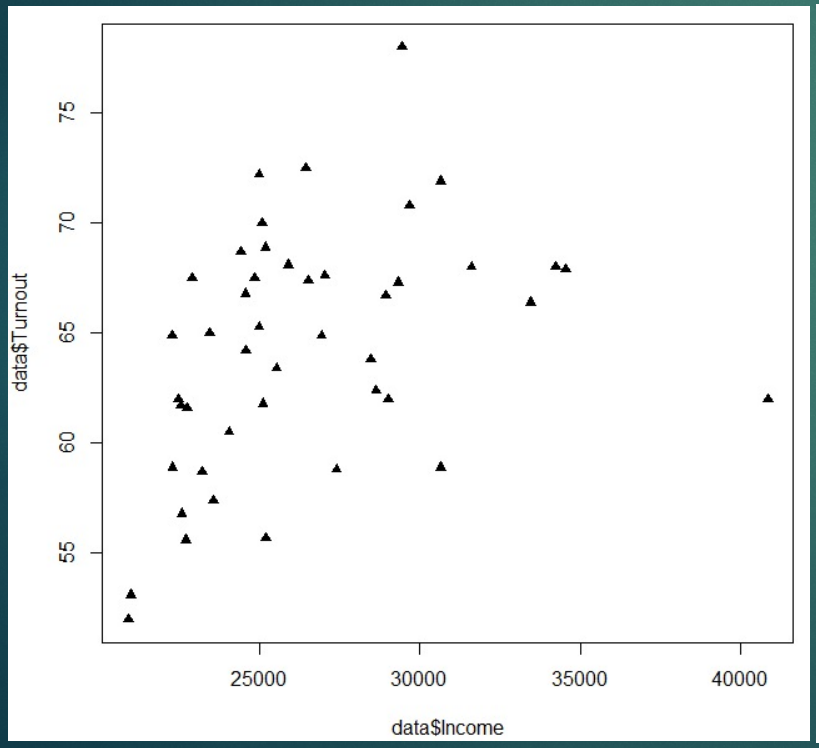
```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> par(bg="blue", col="red")
> plot(mtcars$cyl, mtcars$mpg, bg="white", col="red")
> png("plot.png", width=360, height=360, units="px", pointsize=72)
>
```

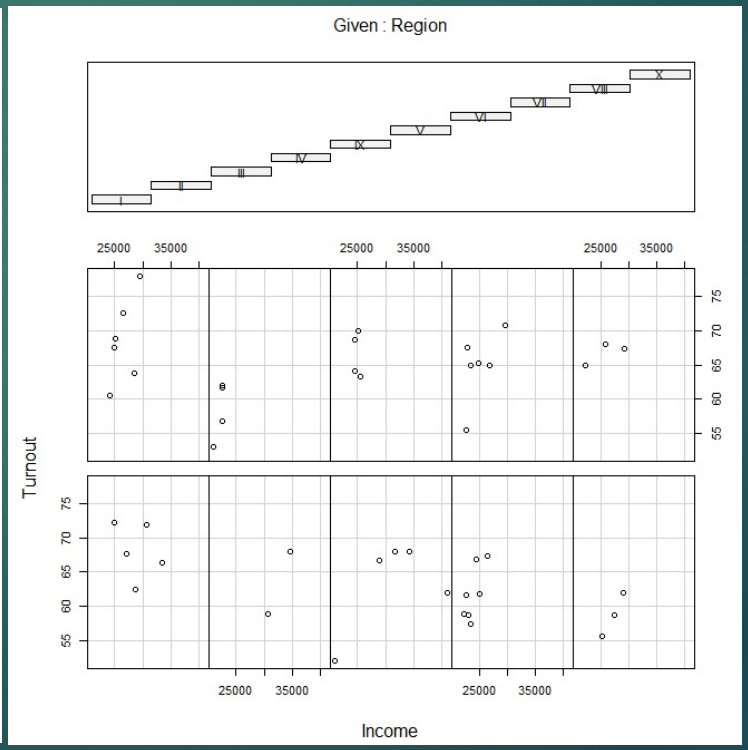


2.2 课堂测试5

3.

```
> data <- read.csv("obama_vs_mccain.csv", header=TRUE)
> data <- na.omit(data)
> plot(data$Income, data$Turnout, pch=17)
>
> coplot(Turnout ~ Income | Region, data=data, column=5)
>
```





2. 课堂测试讲解

ggplot2 II 课堂测试06 40分钟 先用电脑完成 然后誊抄纸上

- 1、查看数据集mtcars，根据要求作图：

  (1)分别使用qplot、ggplot函数画出mpg和wt关系的散点图；

  (2)使用三种方式画出mpg列的直方图,同时在使用qplot和ggplot时指定每个小圆柱体的宽度是4；

  (3)使用三种绘图函数画出mpg变量的密度曲线。

- 2、使用datasets包中的数据集pressure，查看其数据并按要求画图：

  (1)请画出pressure和temperature关系的曲线图；

  (2)分别使用qplot和ggplot画出pressure和temperature关系的散点图和折线图。

- 3、使用datasets中的数据集ToothGrowth，完成如下的绘图要求：

  (1)以supp变量作为分类,分别使用三种绘图函数画出len变量的箱型图。

- 4、使用ggplot2包中数据集mpg，完成练习：

  (1)使用mpg数据集定义一个 ggplot对象，表示hwy与cty的关系；

  (2)画一个散点图，指定颜色有year列来指定，并在上边绘图的基础上画出平滑的拟合曲线；

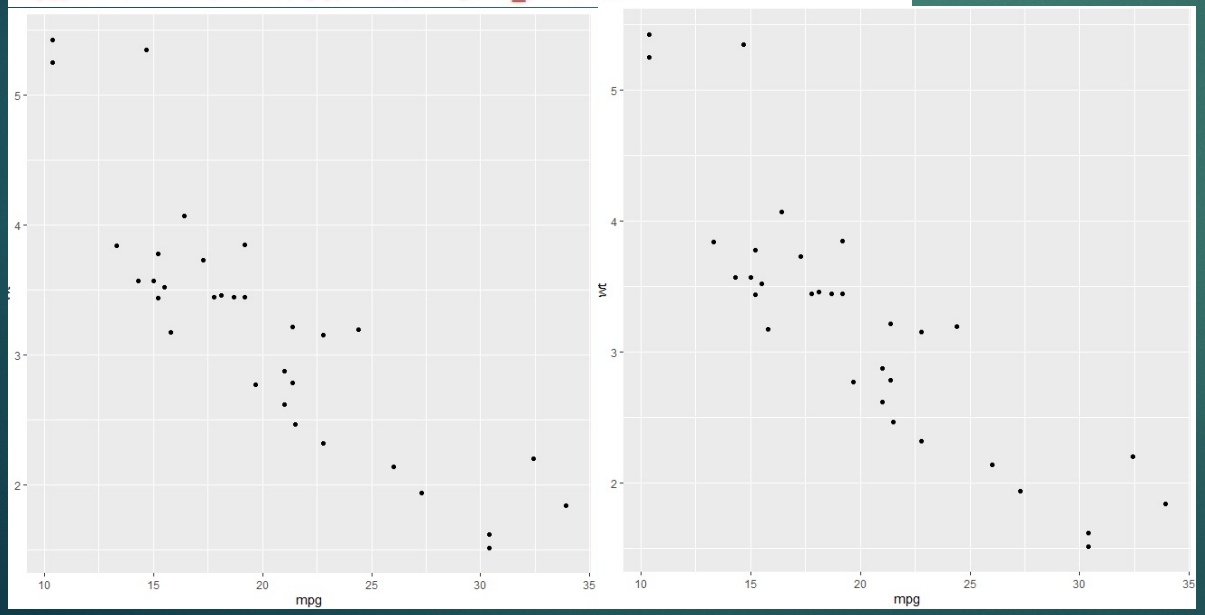
  (3)继续使用(1)中定义的ggplot对象画散点图，使用class来指定颜色，displ指定大小，透明度；
  指定为0.5,position指定为抖动，在散点图的基础上添加拟合曲线；

  (4)使用qplot画出hwy与cty的关系的散点图，并根据year变量分面，同时添加拟合曲线。

2.3 课堂测试6

1.

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> qplot(mpg, wt, data=mtcars)
> ggplot(mtcars, aes(mpg, wt)) + geom_point()
```

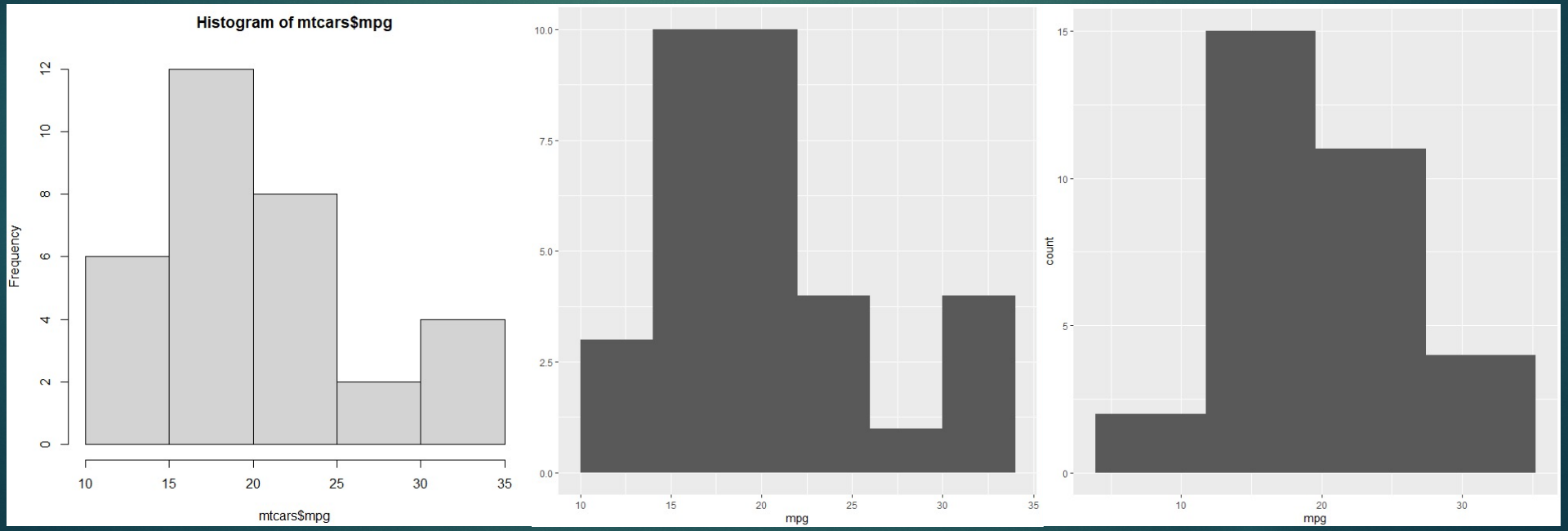


2.3 课堂测试6

1.

```
> hist(mtcars$mpg)
> qplot(mpg, data=mtcars,geom="histogram", binwidth=4)
> ggplot(mtcars, aes(x=mpg)) + geom_histogram(bins=4)
```

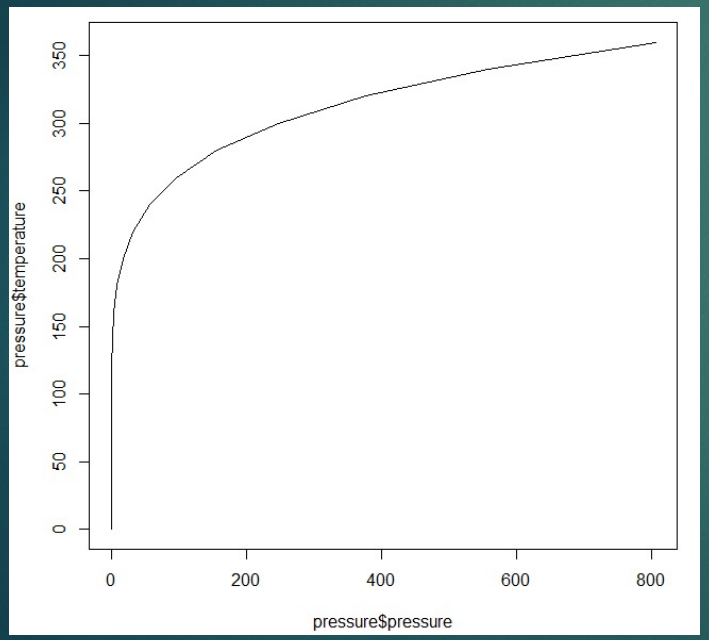
```
> plot(density(mtcars$mpg))
> qplot(mpg, data=mtcars, geom="density")
> ggplot(mtcars, aes(x=mpg)) + geom_density()
>
```

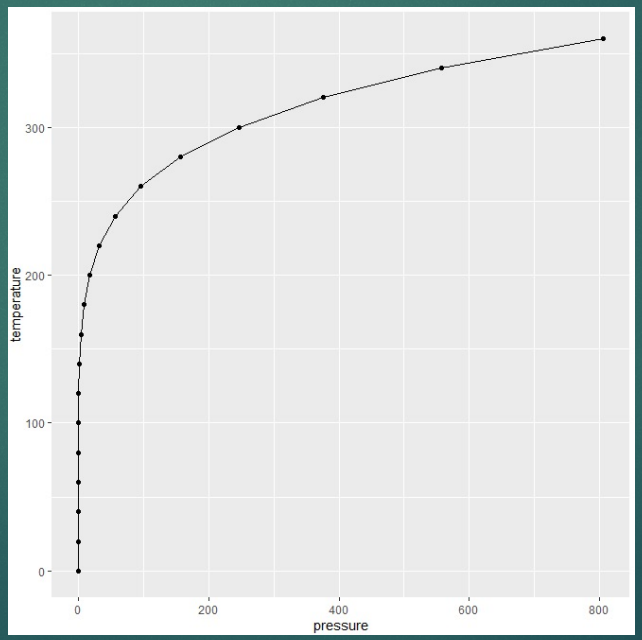


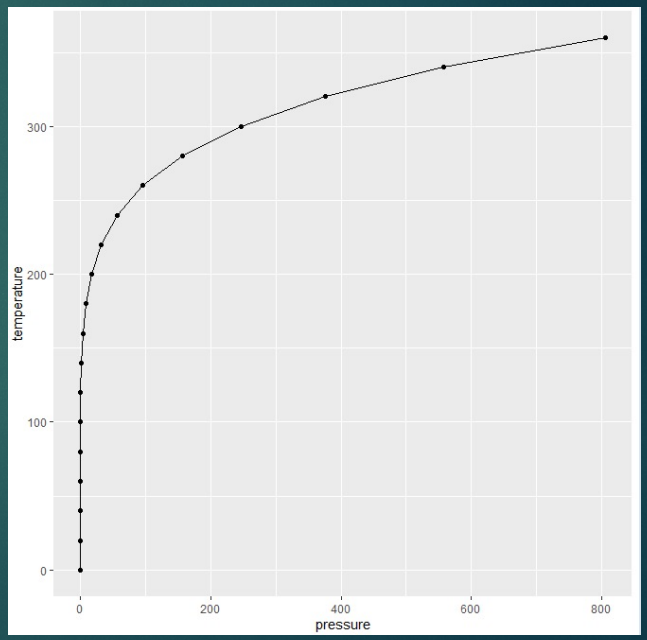
2.3 课堂测试6

2.

```
> str(pressure)
'data.frame':   19 obs. of  2 variables:
 $ temperature: num  0 20 40 60 80 100 120 140 160 180 ...
 $ pressure   : num  0.0002 0.0012 0.006 0.03 0.09 0.27 0.75 1.85 4.2 8.8 ...
> plot(pressure$pressure, pressure$temperature, type="l")
> qplot(pressure, temperature, data=pressure, geom=c("point", "line"))
> ggplot(pressure, aes(pressure, temperature)) + geom_line() + geom_point()
>
```



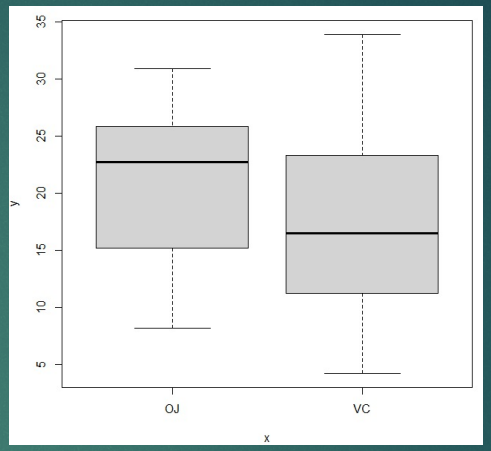


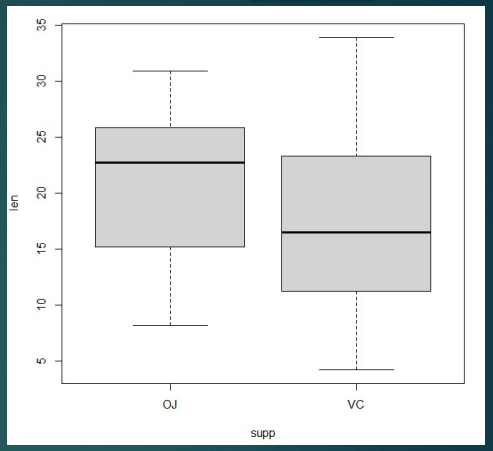


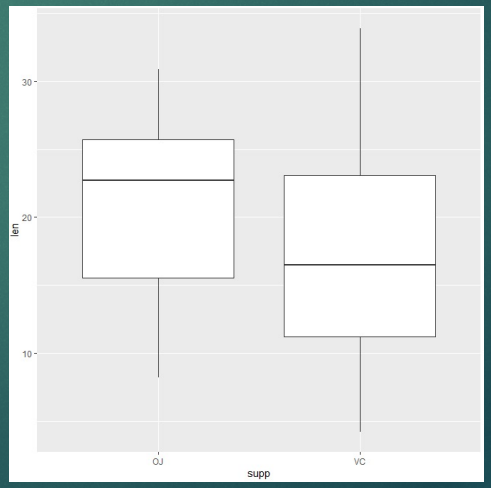
2.3 课堂测试6

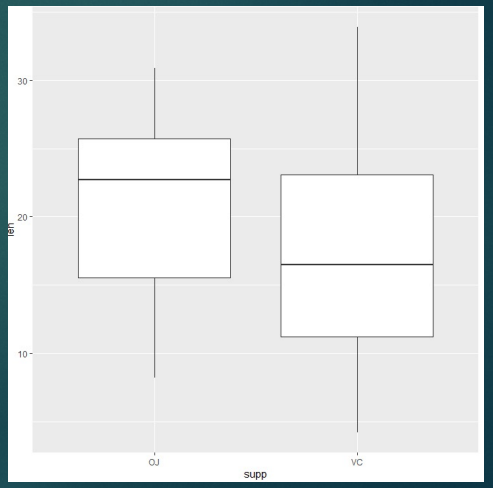
3.

```
> str(ToothGrowth)
'data.frame':   60 obs. of  3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
> plot(ToothGrowth$supp, ToothGrowth$len)
> boxplot(len ~ supp, data=ToothGrowth)
> qplot(supp, len, data=ToothGrowth, geom="boxplot")
> ggplot(ToothGrowth, aes(supp, len)) + geom_boxplot()
> boxplot(supp ~ len, data=ToothGrowth)
```





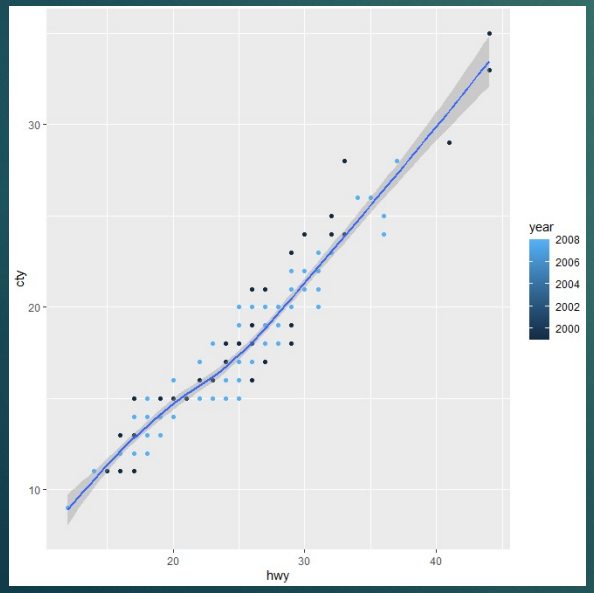


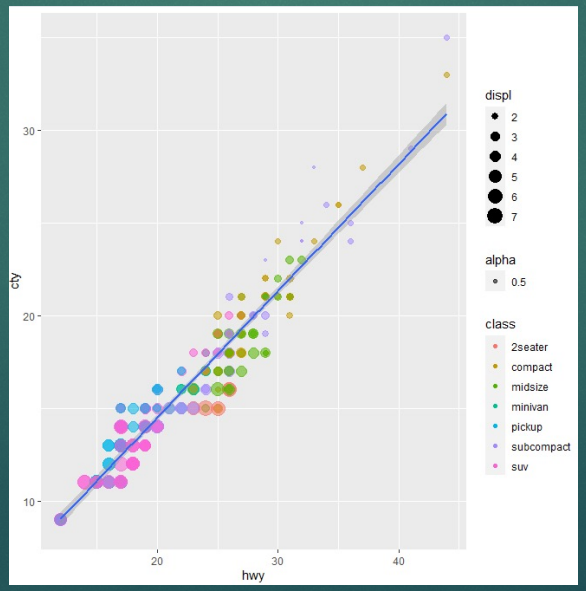


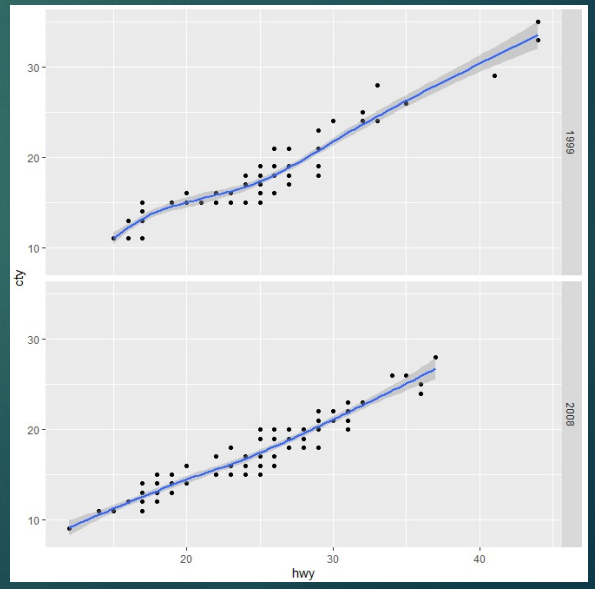
2.3 课堂测试6

4.

```
> str(mpg)
tibble[,11] [234 x 11] (S3: tbl_df/tbl/data.frame)
 $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
 $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
 $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
 $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
 $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
 $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
 $ drv         : chr [1:234] "f" "f" "f" "f" ...
 $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
 $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
 $ fl          : chr [1:234] "p" "p" "p" "p" ...
 $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
> p <- ggplot(mpg, aes(hwy, cty))
> p + geom_point(aes(colour=year)) + geom_smooth()
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
> p + geom_point(aes(colour=class, size=displ,alpha=0.5, position="jitter")) + geom_smooth(method="lm")
`geom_smooth()` using formula 'y ~ x'
Warning message:
Ignoring unknown aesthetics: position
> qplot(hwy, cty, data=mpg, facets=year~., geom=c("point", "smooth"))
```







2. 课堂测试讲解

RGCook 课堂测试07 40分钟 先用电脑完成 然后誊抄纸上

- 使用ggplot2里的画图函数完成以下的练习：
  - 1、将数据集Big_Mart_Dataset.csv,加载到R空间，将数据框命名为mart,查看mart的维度和基本结构。
  - 2、画Item_MRP和Item_Visibility的关系图，要求: (1)指定颜色属性为Item_Type; (2)设置x轴的标度(scale)，x轴名字为Item Visibility", x轴刻度为0-0.35以0.05为间隔的数值序列；设置y轴的标度(scale)，y轴名字为Item MRP , y轴刻度为0-270以30为间隔的数值序列; (3)设置图形主题为theme_bw，图形标题为Scatterplot。
  - 3、在2基础上，根据因子类型的列Item_Type进行分面。
  - 4、画列变量Item_MRP的直方图，要求：(1)每个小圆柱体的宽度为2, (2)设置x轴的标度(scale)，x轴名字为Item MRP, x轴刻度为0-270以30为间隔的数值序列；设置y轴的标度(scale)，y轴名字为Count,y轴刻度为0-200以20为间隔的数值序列;(3)设置标题为"Histogram"

- 使用ggplot2里的画图函数完成以下的练习：
  - 5、画出列变量Outlet_Establishment_Year的条形图，要求(1): 填充色为"red"; (2): 主题为theme_bw和theme_gray;(3): 设置x轴的标度(scale)，x轴名字为Establishment_Year, x轴刻度为1985-2010为间隔的数值序列；设置y轴的标度(scale)，y轴名字为Count , y轴刻度为0-1500以150为间隔的数值序列; (4): 设置标题为Bar Chart，翻转坐标轴
  - 6、画出Outlet_Location_Type堆叠的条形图 (1): 使用 Outlet_Type设置填充色; (2): 设置图形的标题为Stacked Bar Chart，x轴的名称为Outlet Location Type", y轴的名称为Count of Outlets
  - 7、画Outlet_Identifier以Item_Outlet_Sales为分类变量的箱型图;(1): 填充色为红色; (2): y轴名称为"Item Outlet Sales", 坐标为0-15000以150为间隔的数值序列; (3): 设置标题为"Box Plot", x 轴坐标为"Outlet Identifier
  - 8、画列变量Item_Outlet_Sales面积图表 要求: (1)统计变换为 "bin", bin的宽度为30, 填充色为"steelblue";(2)x轴的标度为0-11000以1000间隔的数值序列;(3)图形标题为"Area Chart", x 轴命名为 "Item Outlet Sales", y轴命名为 "Count"。

2.4 课堂测试7

3. 作业

完成DataCamp课程 Market Basket Analysis in R