统计学（R实现）

第一次上机课

赵震宇（2120253538）
南开大学 国际经济研究所
zzynankai@outlook.com
xishanyu2.github.io
2025 年 10 月 16 日

参考资料

- 贾俊平著. 统计学:基于R. 中国人民大学出版社. 2023
- 今井耕介著. 量化社会科学导论. 上海财经大学出版社有限公司. 2020
- R for Data Science (2e): 英文版, 中文版
- 徐嘉烨老师课程网站

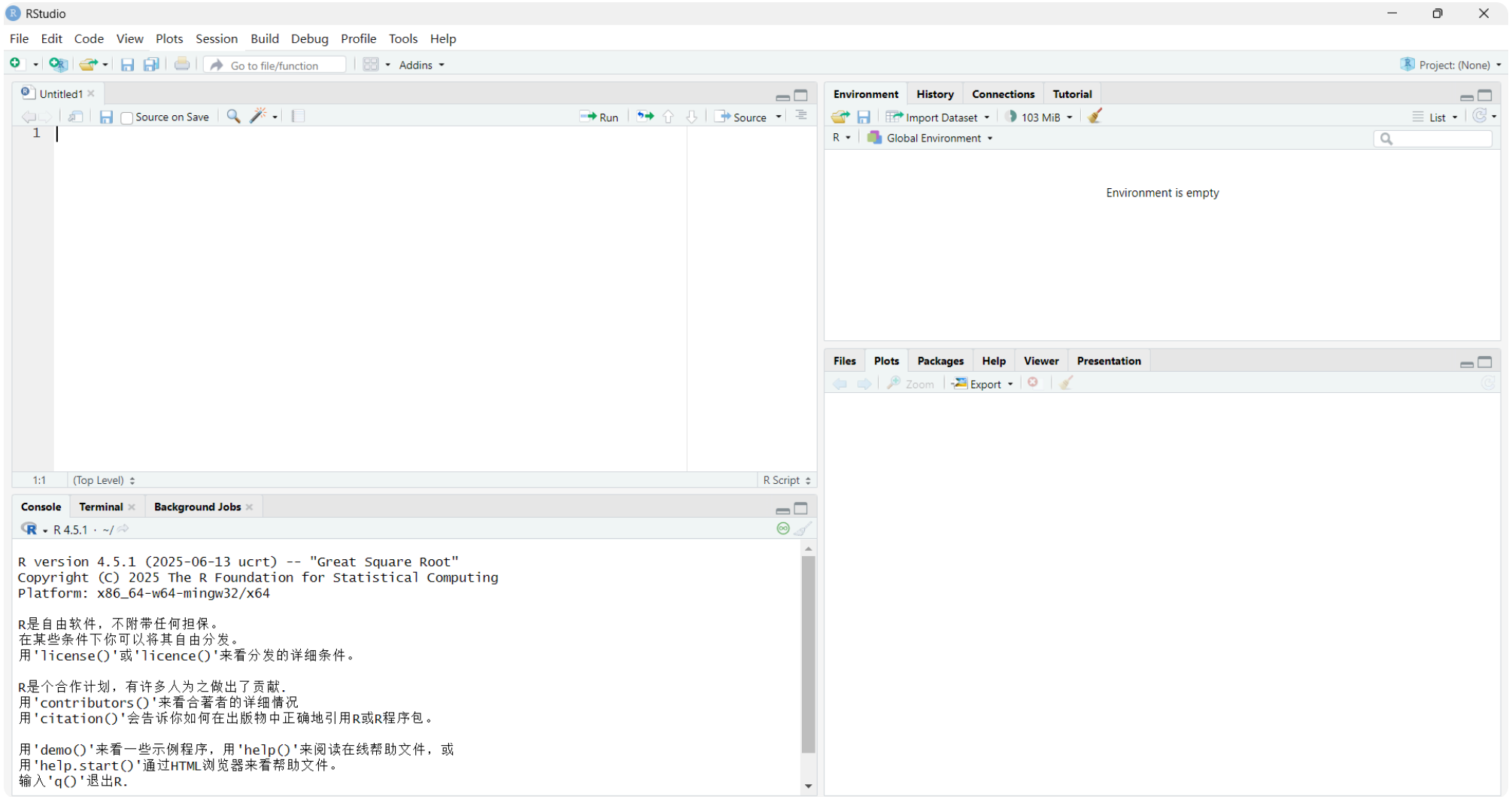
- 第1章 导论
  - 1.1 统计及其应用领域
  - 1.2 统计数据的类型
  - 1.3 统计中的几个基本概念
- 第2章 数据的搜集
  - 2.1 数据的来源
  - 2.2 调查方法
  - 2.3 实验方法
  - 2.4 数据的误差
- 第3章 数据的图表展示
  - 3.1 数据的预处理
  - 3.2 品质数据的整理与展示
  - 3.3 数值型数据的整理与展示
  - 3.4 合理使用图表
- 第4章 数据的概括性度量
  - 4.1 集中趋势的度量
  - 4.2 离散程度的度量
  - 4.3 偏态与峰态的度量
- 第5章 概率与概率分布
  - 5.1 随机事件及其概率
  - 5.2 离散型随机变量及其分布
  - 5.3 连续型随机变量的概率分布
- 第6章 统计量及其抽样分布
  - 6.1 统计量
  - 6.2 由正态分布导出的几个重要分布
  - 6.3 样本均值的分布与中心极限定理
- 第7章 参数估计
  - 7.1 参数估计的基本原理
  - 7.2 一个总体参数的区间估计
  - 7.3 两个总体参数的区间估计
  - 7.4 样本量的确定
- 第8章 假设检验
  - 8.1 假设检验的基本问题
  - 8.2 一个总体参数的检验
  - 8.3 两个总体参数的检验
  - 8.4 检验问题的进一步说明
- 第9章 分类数据分析
  - 9.1 分类数据与统计量
  - 9.2 拟合优度检验
  - 9.3 列联分析：独立性检验
  - 9.4 列联表中的相关测量
  - 9.5 列联分析中应注意的问题
- 第10章 方差分析
  - 10.1 方差分析引论
  - 10.2 单因素方差分析
  - 10.3 双因素方差分析
- 第11章 一元线性回归
  - 11.1 变量间关系的度量
  - 11.2 一元线性回归
  - 11.3 利用回归方程进行预测
  - 11.4 残差分析
- 第12章 多元线性回归
  - 12.1 多元线性回归模型
  - 12.2 回归方程的拟合优度
  - 12.3 显著性检验
  - 12.4 多重共线性
  - 12.5 利用回归方程进行预测
  - 12.6 变量选择与逐步回归
- 第13章 时间序列分析和预测
  - 13.1 时间序列及其分解
  - 13.2 时间序列的描述性分析
  - 13.3 时间序列预测的程序
  - 13.4 平稳序列的预测
  - 13.5 趋势型序列的预测
  - 13.6 复合型序列的分解预测
- 第14章 指数
  - 14.1 基本问题
  - 14.2 总指数编制方法
  - 14.3 指数体系
  - 14.4 几种典型的指数
  - 14.5 综合评价指数

CONTENT

目录

1. R基础

1.1 RStudio


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment History Connections Tutorial
Import Dataset
103 MiB
List
R
Global Environment
Environment is empty
Files Plots Packages Help Viewer Presentation
Zoom
Export
Console Terminal Background Jobs
R 4.5.1 · ~/
R version 4.5.1 (2025-06-13 ucrt) -- "Great Square Root"
Copyright (C) 2025 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64
R是自由软件，不附带任何担保。
在某些条件下你可以将其自由分发。
用'license()'或'licence()'来看分发的详细条件。
R是个合作计划，有许多人为之做出了贡献.
用'contributors()'来看合著者的详细情况
用'citation()'会告诉你如何在出版物中正确地引用R或R程序包。
用'demo()'来看一些示例程序，用'help()'来阅读在线帮助文件，或
用'help.start()'通过HTML浏览器来看帮助文件。
输入'q()'退出R.

1.2 R初识

```
5 + 3
sqrt(4)
class(sqrt)

result <- 5+3
result
class(result)

zzy <- "teaching_assistant"
class(zzy)
```

```
x<-c(80,87,98,73,100)
sum(x)
?sum # help(sum)
install.packages("rio")
library(rio)
```

```r
# 向量
c1<-c(2,3,4,5)
c2<-c(6,7,8)
c3<-c(c1,c2)

# 索引
c1[2]
c1[c(2,4)] # 3 5
c1[(4,2)] # 5 3
c1[-3] #移除对应元素 2 3 5
c4<-c1*100
c1[c(1,2)]<-c(1,9) # 1 9 3 4 5
```

```r
# 函数
length(c1)
max(c1)
range(c1)

1:10
year<-seq(from=2010,to=2025,by=5) # by（步长）/length（个数）
names(c1)<-year
```

```r
myfunction<-function(input){
        DEFINE "output" USING INPUTS
        return(output)
}
```

```
# 文件路径
getwd #working directory
setwd()

# 读取数据
table1_1<-read.csv("F:/rdata/example/chap01/table1_1.csv")
class(table1_1) # "data.frame"
names(table1_1)
nrow(table1_1)
ncol(table1_1)
dim(table1_1)
summary(table1_1) # str(table1_1)

read.dta("auto.dta") # install.packages("foreign")
read.spss("ex3_1.sav")

# 保存数据

write.dta(table1_1,file="table1_1.dta")
```

1.3 R数据类型

- 向量（vector）
- 矩阵（matrix）
- 数组（array）
- 数据框（data frame）
- 因子（factor）
- 列表（list）

1.3.1 向量

```
a<-c(1,2,3)
b<-rep(1:3,times=3)
c<-rep(1:3,each=3)
rev(a) # reverse 3 2 1
LETTERS[1:10]
letters[1:10]
month.abb #缩写的月份
month.name[1:6]
```

1.3.2 矩阵

```
v<-1:6
mat<-matrix(v,nrow=2,ncol=3,byrow=TRUE) #默认FALSE按列填充
rownames(mat)=c("甲","乙")
colnames(mat)=c("A","B","C")
t(mat) #转置
```

```
x <- c(1,2,3,4)
x %*% t(x) # (4 by 1) times (1 by 4)
t(x) %*% x # (1 by 4) times (4 by 1)
```

1.3.3 数据框

```
# 查看数据框
names<-c("甲","乙","丙","丁","戊")
stat<-c(68,85,74,88,63)
math<-c(85,91,74,100,82)
econ<-c(84,63,61,49,89)
table2_1<-data.frame(姓名=names,统计学=stat,数学=math,经济学=econ)

# 访问数据框
table2_1$统计学
table2_1[,2:3] #前行后列，等价于table2_1[,c(2,3)]
table2_1[3,2]

# 数据框合并与排序
table_merge<-rbind(table1,table2)
arrange(table2_1,姓名)
arrange(table2_1,desc(数学))
```

1.3.4 因子

```r
a<-c("金融","地产","医药","医药","金融","医药")
f1<-factor(a);f1 # Levels: 地产 金融 医药
as.numeric(f1) # 2 1 3 3 2 3

# 无序因子转换为有序因子和数值
b<-c("优","良","中","差")
f2<-factor(b,ordered=TRUE,levels=c("优","良","中","差"))
f2 # Levels: 优 < 良 < 中 < 差
as.numeric(f2) # 1 2 3 4
```

2. 数据可视化

Data visualization

> “The simple graph has brought more information to the data analyst’s mind than any other device.” — John Tukey

R for data visualization

- R语言具有强大的可视化功能，可绘制样式繁多的图形，其中包括base和grid两大底层绘图系统。
- 安装R时自带graphics绘图包；同一种图形可以使用不同的包来实现。
- ggplot2是最优雅和最多功能的之一（也是相对容易上手的），以下多数图形基于grid开发的ggplot包来实现。

拓展阅读

- 可乐-CSDN博客

2.1 使用 ggplot2 可视化数据

```
# R调包
install.packages("tidyverse")
library(tidyverse) # ggplot2是tidyverse中的核心包之一
install.packages('ggthemes', dependencies = TRUE)
library(ggthemes)
```

2.1.1 penguins data frame

相关关系？因果关系？

人的身高和体重间有相关关系吗？车的长度和宽度间呢？
长脚蹼的企鹅比短脚蹼的企鹅体重更重还是更轻呢？
脚蹼长度和体重之间的关系是什么样的？是正相关还是负相关？是线性还是非线性的？这种关系是否因企鹅的物种而异？岛屿的差异是否会对这种关系产生影响？让我们创建可视化图表来回答这些问题。

2.1.1 penguins data frame

- penguins数据集包含了Palmer群岛三个岛屿上企鹅的身体测量数据
- penguins包含344个观测值

回顾统计学第一章所学：变量、数据

变量（variable）：所有企鹅的属性。
值（value）：测量时variable所处的状态。
观测（observation）：单个企鹅的所有属性。一个observation包含多个values
Questions：总体？样本？样本量？
表格数据（Tabular data）：value——在“单元格”中，variable——列，observation——行

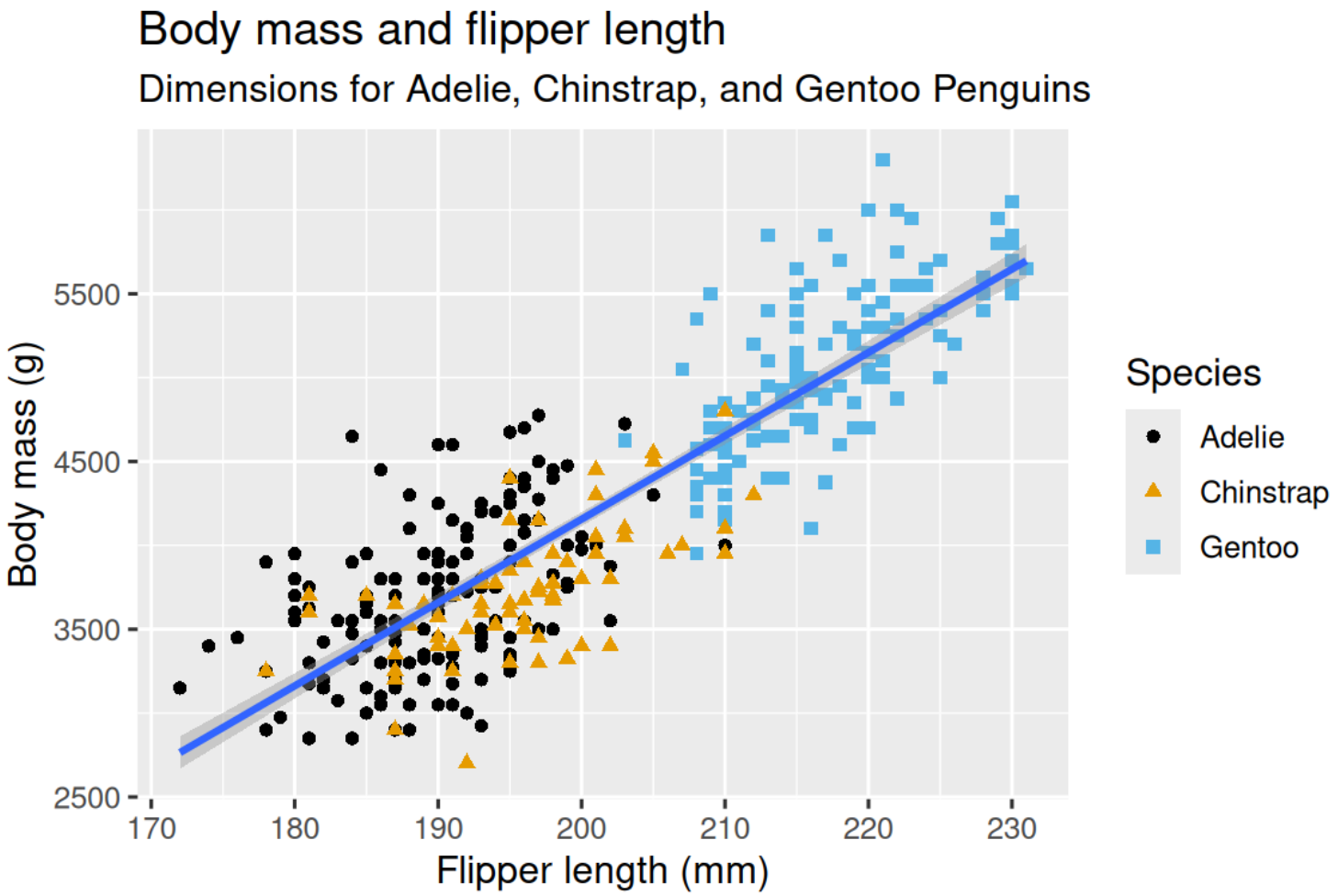
2.1.1 penguins data frame

```
penguins # penguins数据集已内置在datasets包中，无需额外安装palmerpenguins
glimpse(penguins) # 转置?
head(penguins,5)
```

```
  species    island    bill_len bill_dep flipper_len body_mass sex    year
1  Adelie Torgersen      39.1     18.7          181      3750   male 2007
2  Adelie Torgersen      39.5     17.4          186      3800 female 2007
3  Adelie Torgersen      40.3     18.0          195      3250 female 2007
4  Adelie Torgersen        NA       NA           NA        NA   <NA> 2007
5  Adelie Torgersen      36.7     19.3          193      3450 female 2007
```

1. `species`: a penguin’s species (Adelie, Chinstrap, or Gentoo).
2. `flipper_len`: length of a penguin’s flipper, in millimeters.
3. `body_mass`: body mass of a penguin, in grams.

2.1.2 可视化——在R里加图层！

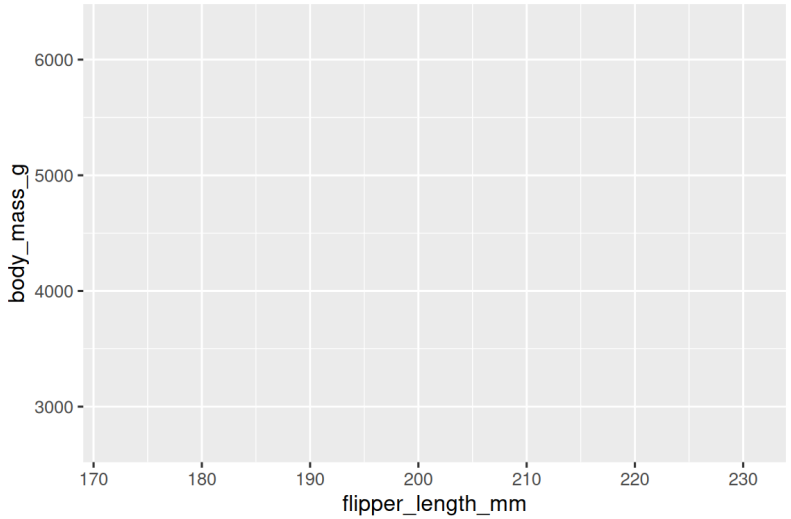


Step1：使用函数 ggplot() 开始绘图

```
ggplot(data = penguins) # 创建一个空图表（empty graph）
```

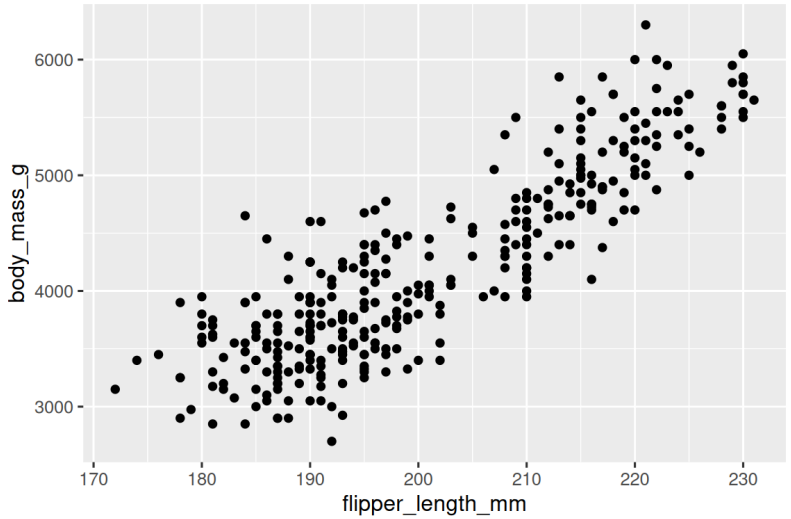
Step2：使用参数 mapping 将数据映射到图表

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass_g)
)
```



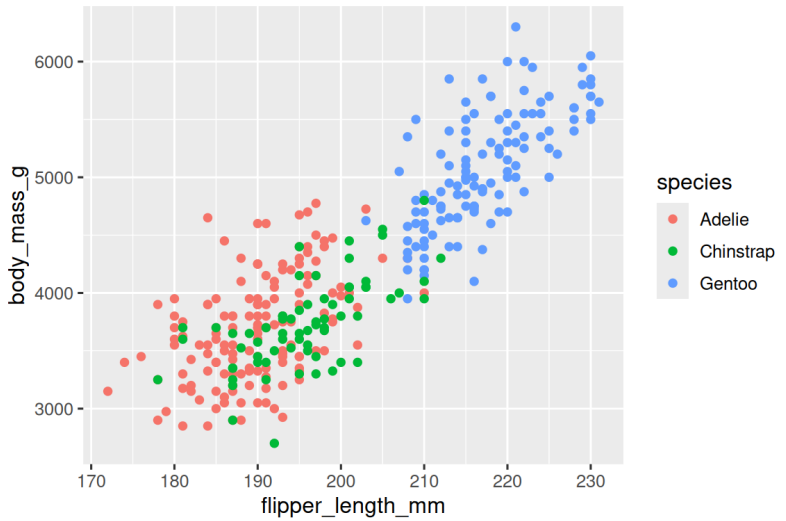
Step3：使用 geom_point() 创建散点图

```
ggplot(
  data = penguins, # 可不写data =
  mapping = aes(x = flipper_len, y = body_mass_g) # 可不写mapping =
)+
geom_point()
```



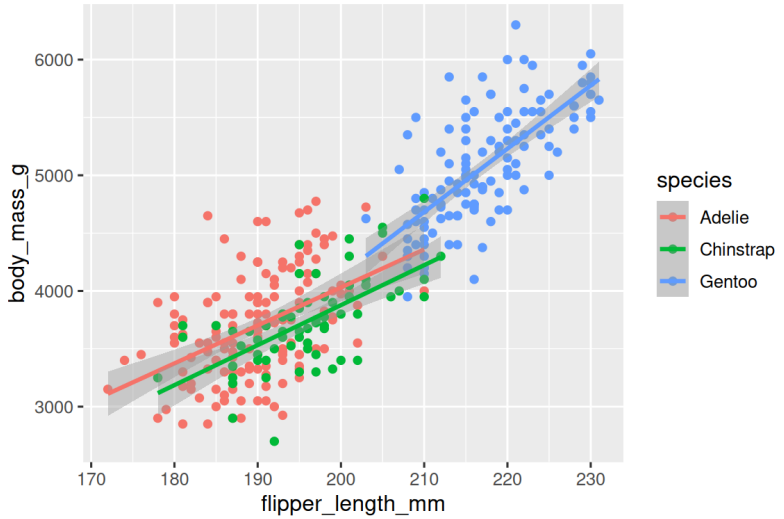
Step4：使用不同颜色的点来表示不同物种

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass_g, color = species)
)+
geom_point()
```



Step5：使用线性模型（linear model）绘制最佳拟合线

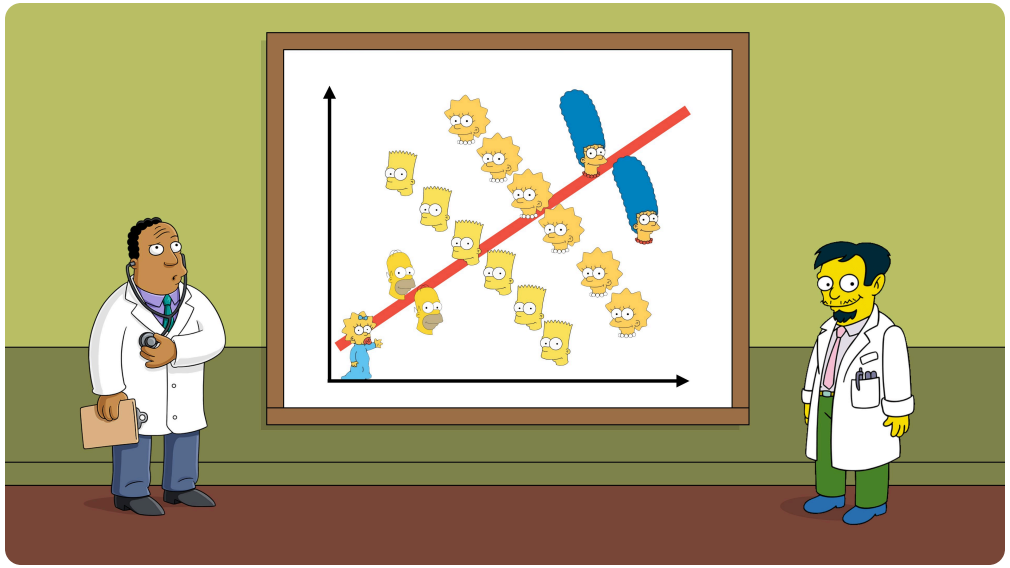
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass_g, color = species)
)+
geom_point()+
geom_smooth(method = "lm")
```



知识拓展

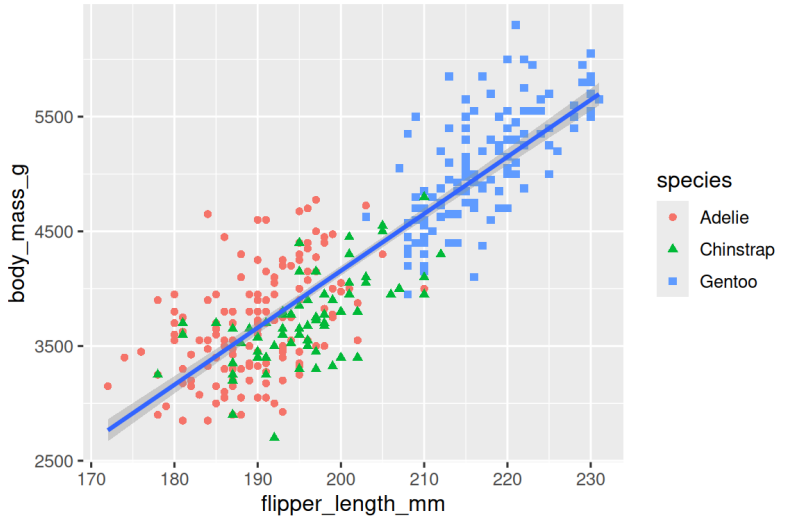
辛普森悖论（Simpson's Paradox）

两个变量X和Y在每个分组中的关系是正/负，但在总体（所有组加总）中关系会发生逆转变成负/正。



Step6：在局部级别（local level）上进行 aesthetic mappings

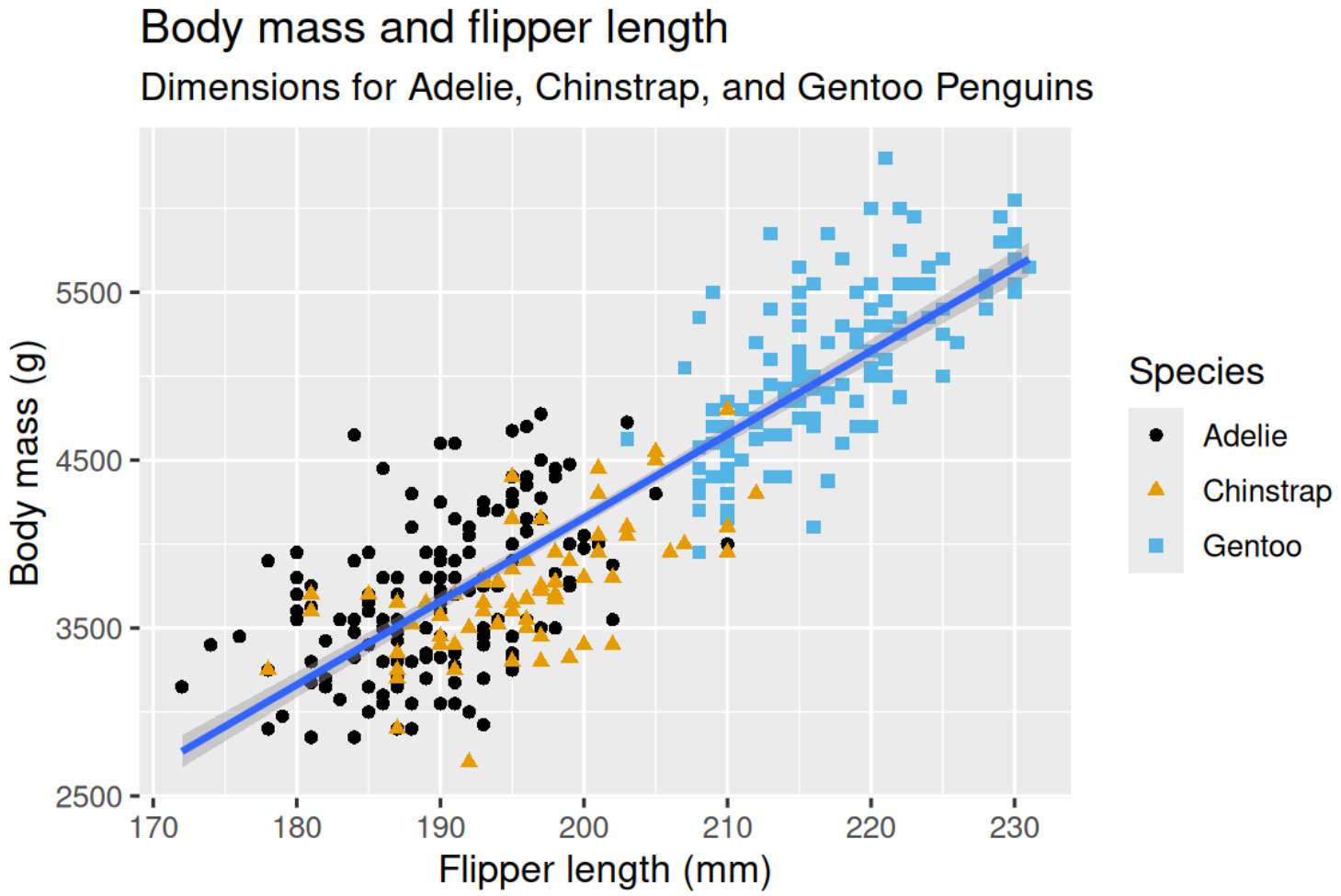
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass_g)
)+
geom_point(mapping = aes(color = species))+
geom_smooth(method = "lm")
```



Step7：添加标签、改变主题

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass)
)+
  geom_point(mapping = aes(color = species, shape = species))+
  geom_smooth(method = "lm")+
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",
    x = "Flipper length (mm)", y = "Body mass (g)",
    color = "Species", shape = "Species"
  ) +
    scale_color_colorblind()
```

We made it!


Body mass and flipper length
Dimensions for Adelie, Chinstrap, and Gentoo Penguins
Body mass (g)
5500
4500
3500
2500
170
180
190
200
210
220
230
Flipper length (mm)
Species
Adelie
Chinstrap
Gentoo

2.1.3 Exercises

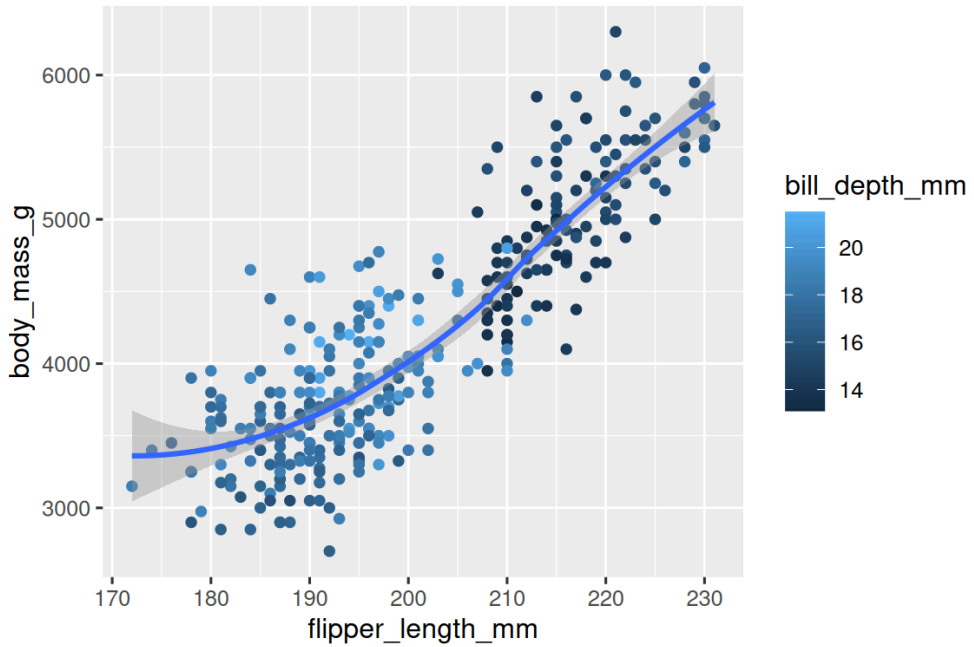
1. `penguins` 有多少行（rows）？ 有多少列（columns）？
2. `penguins` 数据框中的 `bill_dep` 变量描述了什么？
3. 创建一个 `bill_dep` 和 `bill_len` 的散点图。
4. 绘制 `species` 和 `bill_dep` 的散点图，你发现了什么？ 选择什么样的图形可能更好？
5. 为什么下面的代码会出错，如何修改它？

```
ggplot(data = penguins) +
  geom_point()
```

6. 在 `geom_point()` 中，`na.rm` 参数的作用是什么？ 这个参数的默认值是什么？ 创建一个散点图，并将该参数设置为 `TRUE`。

7. 在前面绘制的图中添加以下说明（`caption`）：“Data come from the palmerpenguins package.”

8. 创建以下可视化图形。 `bill_dep` 应该映射到全局级别（global level）还是几何级别（geom level）？



9. 在脑海中运行此代码。 然后在 R 中运行代码检验你的结果。

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass, color = island)
) +
  geom_point() +
  geom_smooth(se = FALSE) # se表示置信区间
```

10. 这两张图看起来会不一样吗？为什么/为什么不？

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_len, y = body_mass)
) +
  geom_point() +
  geom_smooth()
```

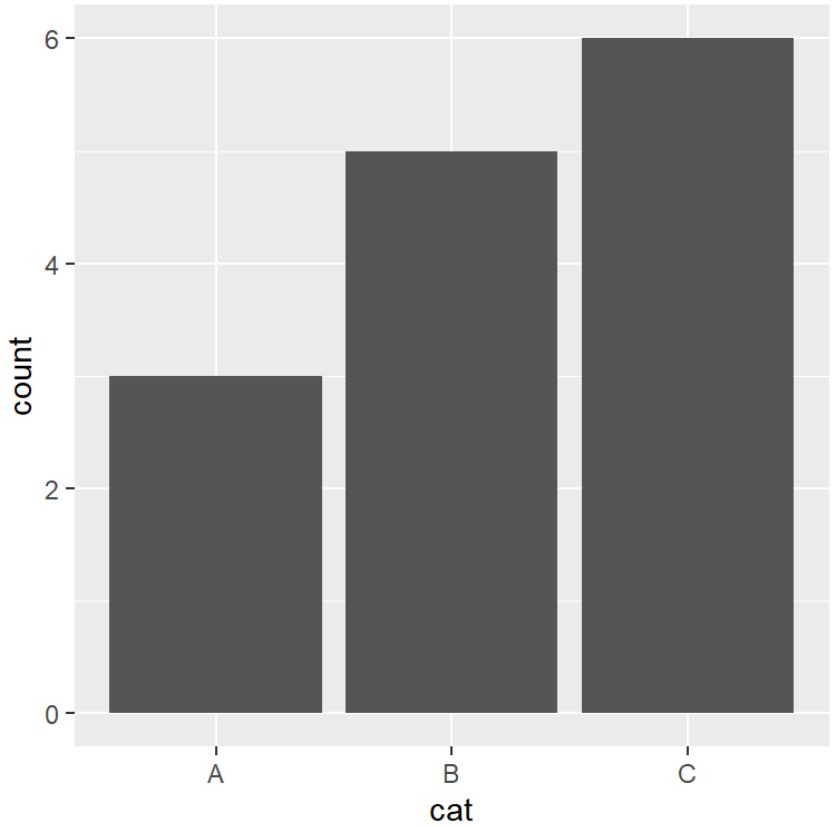
```
ggplot() +
  geom_point(
    data = penguins,
    mapping = aes(x = flipper_len, y = body_mass)
  ) +
  geom_smooth(
    data = penguins,
    mapping = aes(x = flipper_len, y = body_mass)
  )
```

2.2 类别数据可视化

条形图

```
df1 <- data.frame(cat = c("A", "A", "A",
                          "B", "B", "B", "B", "B",
                          "C", "C", "C", "C", "C", "C"))
```

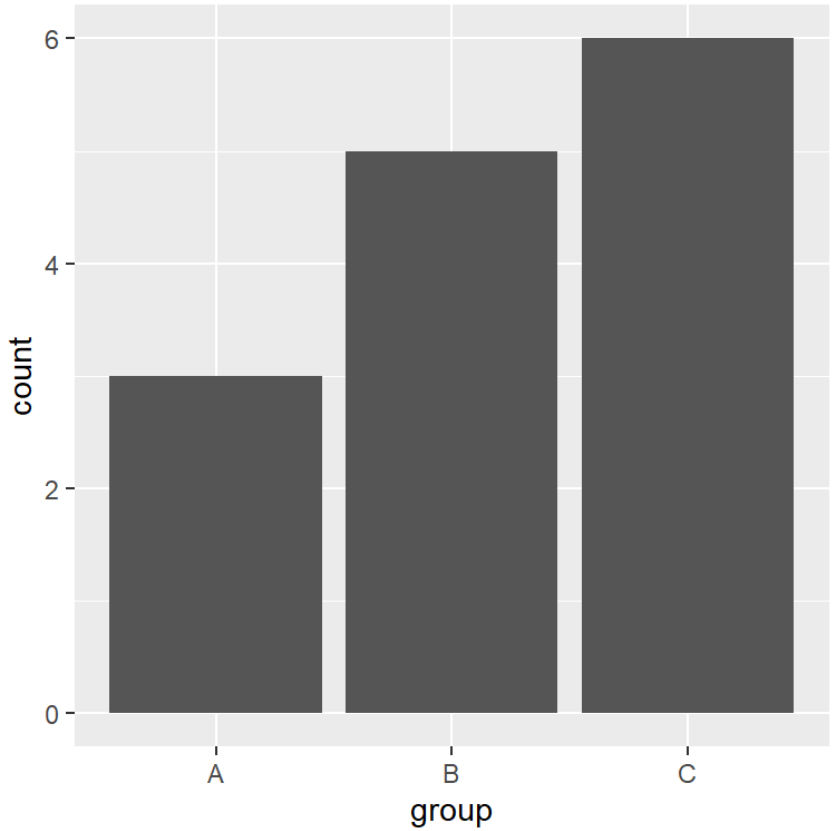
```
ggplot(df1, aes(x = cat)) +
  geom_bar()
```



条形图

```
df2 <- data.frame(group = c("A", "B", "C"),
                  count = c(3, 5, 6))
```

```
ggplot(df2, aes(x = group, y = count)) +
  geom_bar(stat = "identity")
```



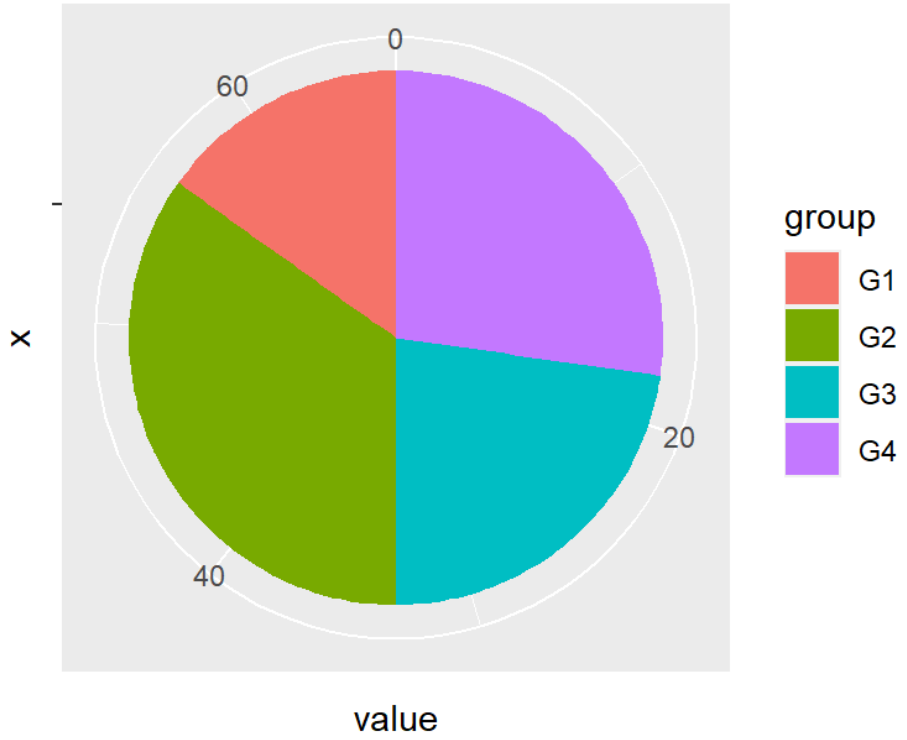
拓展

水平条形图？堆叠条形图？帕累托图？

饼图

```
df <- data.frame(value = c(10, 23, 15, 18),
                 group = paste0("G", 1:4))
```

```
ggplot(df, aes(x = "", y = value, fill = group)) +
  geom_col() +
  coord_polar(theta = "y")
```



拓展

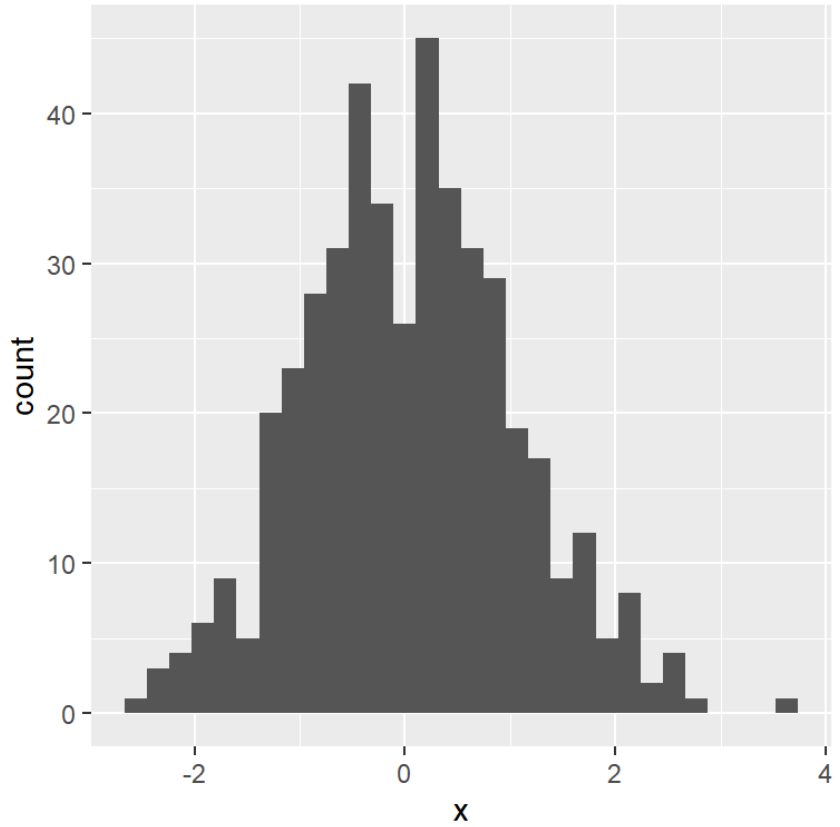
如何添加每一份的数值/百分比标签？环形图？

2.3 数据分布可视化

直方图

```
set.seed(3)
x <- rnorm(450)
df <- data.frame(x)
hist(x)

ggplot(df, aes(x = x)) +
  geom_histogram()
```



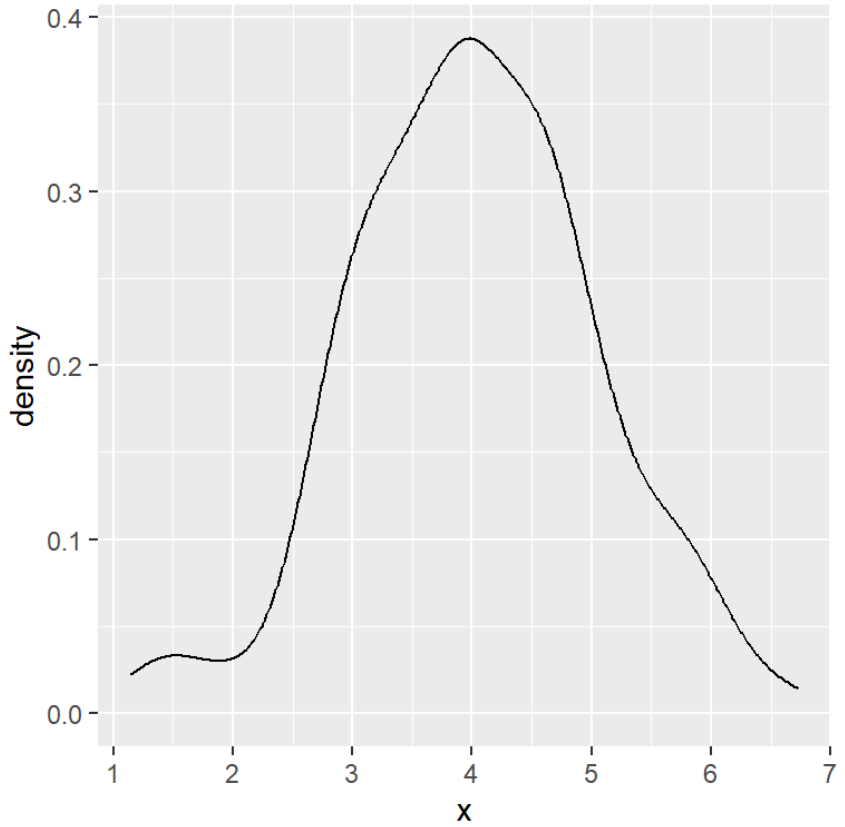
拓展

如何设置bins的数值/数量？

核密度图

```
set.seed(14012021)
x <- rnorm(200, mean = 4)
df <- data.frame(x)

ggplot(df, aes(x = x)) +
  geom_density()
```



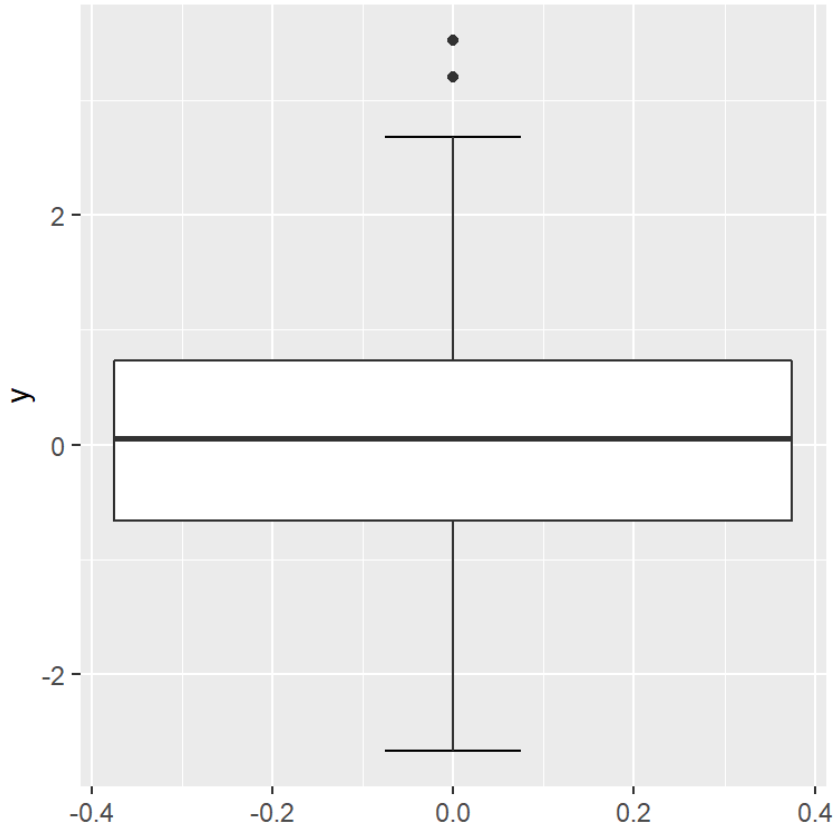
拓展

分组核密度图？脊线图？

作业：直方图堆叠核密度图

箱线图

```
set.seed(3)
y <- rnorm(500)
df <- data.frame(y)
boxplot(y)
ggplot(df, aes(y = y)) +
  stat_boxplot(geom = "errorbar",
               width = 0.15) +
  geom_boxplot()
```



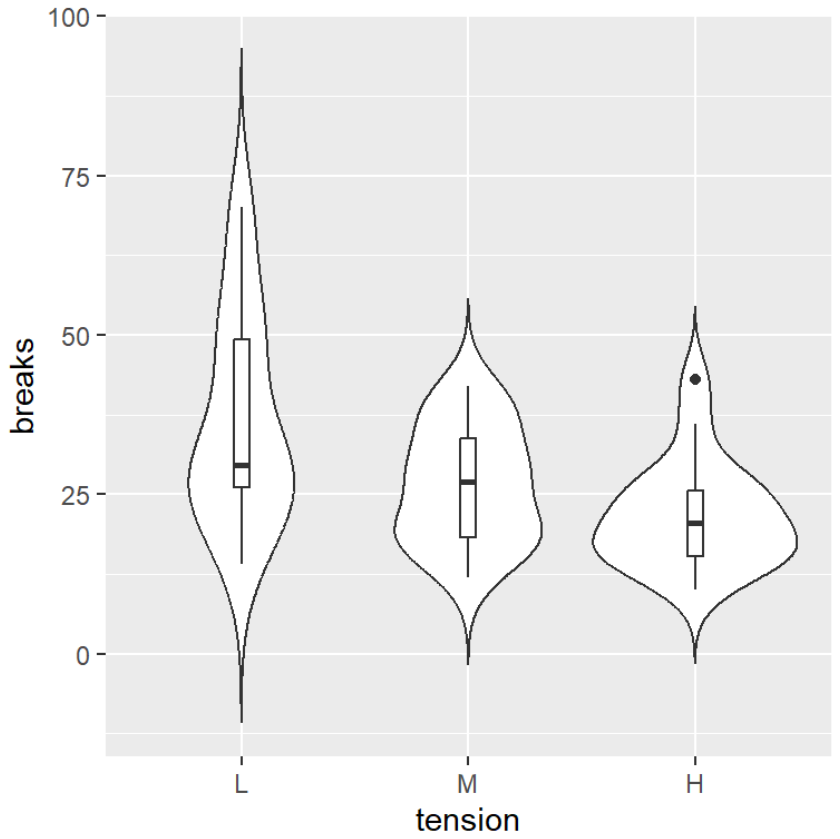
拓展

箱线图与直方图/核密度图的对应关系？

分组箱线图？

小提琴图

```
ggplot(warpbreaks, aes(x = tension, y = breaks)) +
  geom_violin(trim = FALSE) +
  geom_boxplot(width = 0.07)
```



拓展

小提琴图与箱线图的区别与联系？

茎叶图（自学）

- aplpack包

2.4 变量关系可视化

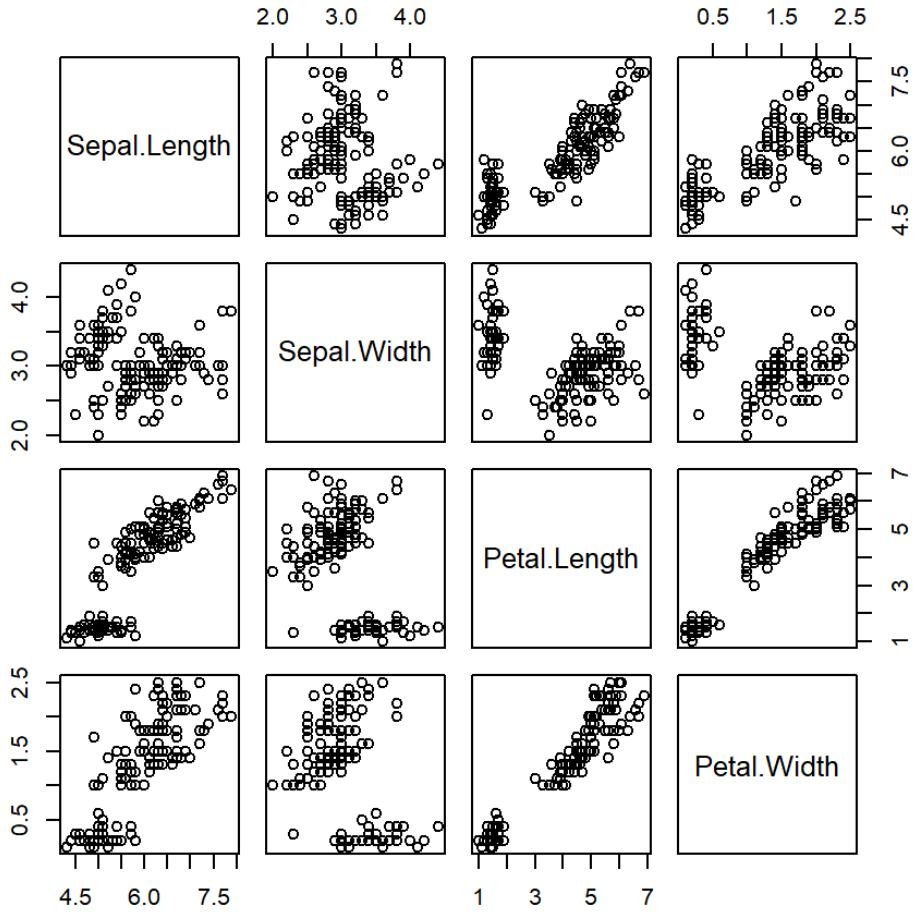
散点图（略）

拓展

边际图为直方图/核密度图的散点图？

散点图矩阵

```
df <- iris[1:4]
pairs(df)
pairs(~ Sepal.Length + Sepal.Width +
        Petal.Length + Petal.Width, data = df)
with(df, pairs(~ Sepal.Length + Sepal.Width +
                 Petal.Length + Petal.Width))
plot(df)
```

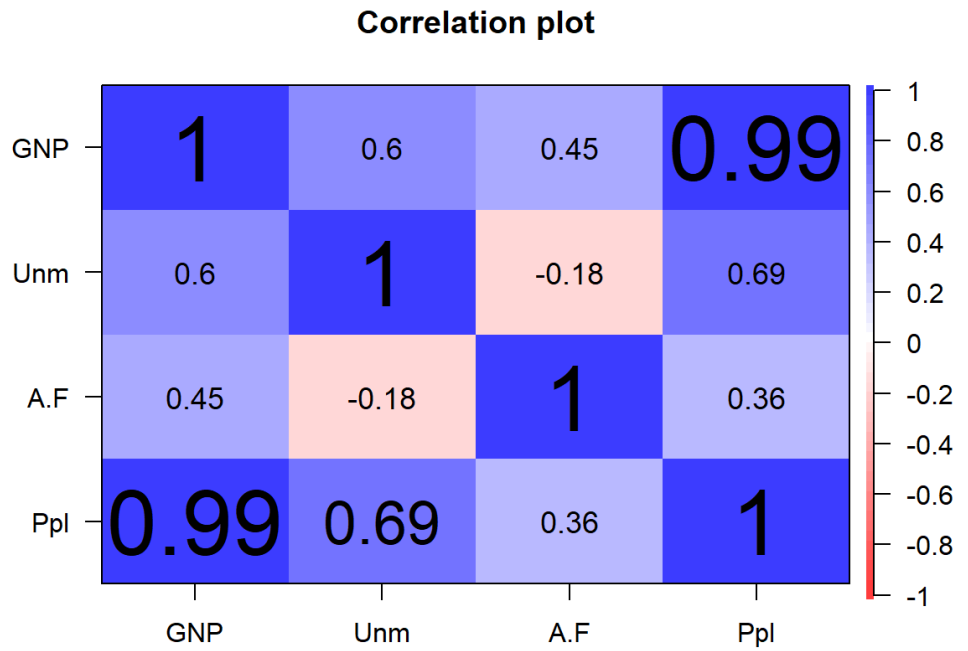


相关系数矩阵

```
install.packages("psych")
library(psych)
corPlot(longley[, 2:5])
```

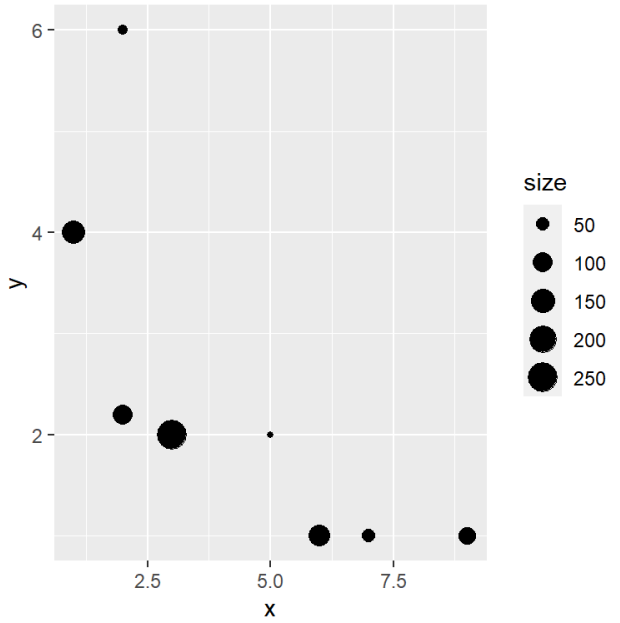
拓展

热图？



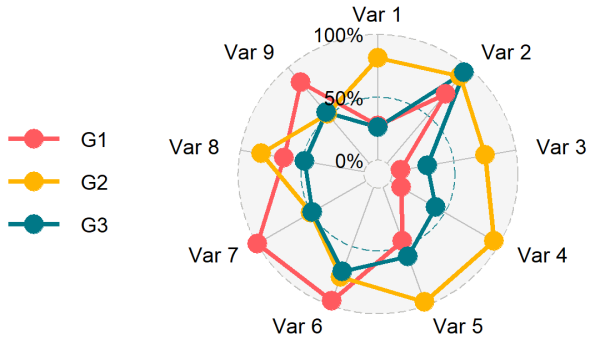
气泡图

```
x <- c(2, 5, 7, 3, 6, 1, 9, 2)
y <- c(2.2, 2, 1, 2, 1, 4, 1, 6)
size <- c(100, 30, 50, 250, 120, 140, 80, 36)
group <- c("A", "A", "A", "B", "C", "B", "D", "B")
df <- data.frame(x = x, y = y, size = size, group = group)
ggplot(df, aes(x = x, y = y, size = size)) +
  geom_point()
```



雷达图

```
install.packages("devtools")
devtools::install_github("ricardo-bion/ggradar")
library(ggradar)
ggradar(df)
```



3. 描述性统计

3.1 描述水平的统计量

简单平均数

```
example4_1<-read.csv("F:/rdata/example/chap04/example4_1.csv")
mean(example4_1$分数)
```

加权平均数

```
example4_2<-read.csv("F:/rdata/example/chap04/example4_2.csv")
weighted.mean(example4_2$组中值,example4_2$人数) # weighted.mean(x,w)
```

中位数

```
median(example4_1$分数)
```

四分位数

```
quantile(example4_1$分数,probs=c(0.25,0.75)) # 默认type=7
```

百分位数

```
quantile(example4_1$分数,probs=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9))
```

众数

```
install.packages("DescTools")
library(DescTools)
Mode(example4_1$分数)
```

平均数、众数与中位数

Questions: example4_1中分数的分布图如何？是左偏还是右偏？

3.2 描述差异的统计量

极差

```
diff(range(example4_1$分数))
```

四分位差

```
IQR(example4_1$分数)
```

方差

```
var(example4_1$分数)
```

标准差

```
sd(example4_1$分数)
```

加权标准差

```
m=example4_2$组中值;f=example4_2$人数
wm<-weighted.mean(m,f)
sqrt(sum((m-wm)^2*f)/(sum(f)-1))
```

变异系数

```
example4_10<-read.csv("F:/rdata/example/chap04/example4_10.csv")
mean1<-mean(example4_10$互联网公司)
mean2<-mean(example4_10$机械制造公司)
sd1<-sd(example4_10$互联网公司)
sd2<-sd(example4_10$机械制造公司)
cv1<-sd1/mean1
cv2<-sd2/mean2
```

```
mean<-apply(example4_10, MARGIN=2, mean) # MARGIN=1按行
sd<-apply(example4_10, MARGIN=2, sd)
cv<-sd/mean
round(cv,4) # 保留四位小数
```

3.3 描述分布形状的统计量

偏度系数

```
library(e1071)
skewness(example4_1$分数)
```

峰度系数

```
kurtosis(example4_1$分数)
```

3.4 数据标准化

标准分数

```
scale(example4_1$分数)
```

极值标准化

```
d<-example4_1$分数
T<-(d-min(d))/(max(d)-min(d))
```

4. 课后习题

习题1（数据可视化）

mtcars是R自带的数据集，该数据集摘自1974年《美国汽车趋势》杂志，包括32款汽车（1973-1974年款）的油耗、汽车设计和性能等共11个变量。根据该数据集绘制以下图形。

(1)绘制气缸数量(cyl)的直方图，并为直方图叠加核密度曲线。

(2)绘制每加仑油行驶的英里数(mpg)和汽车自重(wt)两个变量的箱线图和小提琴图。

(3)绘制该数据集的散点图矩阵和相关系数矩阵图。

(4)绘制每加仑油行驶的英里数(mpg)、总马力(hp)和汽车自重(wt)3个变量的3D散点图和气泡图。

习题2（描述性统计）

随机抽取50个网络购物的消费者，调查他们某月的网购金额（单位:元），结果见exercise4_1.csv。

(1)计算平均数、标准差、极差和四分位差。
(2)计算10%、50%、75%的分位数。
(3)计算偏度系数和峰度系数，分析网购金额的分布特征。
(4)计算标准分数和极值标准化值，检测数据的离群点。

作业要求

- 提交代码+对应图形+必要解释，注意标明题号
- 整理成一个pdf文档，命名为学号-姓名如“2410000-张三”
- **DDL**：10月23日24点前发送至助教飞书（2120253538）
- 视代码完整性、作图美观性、答案准确性评分
- 逾期提交、抄袭雷同记0分
- 可以使用AI

欢迎交流~

zzynankai@outlook.com

Bilibili: 西山yu

xishanyu2.github.io