实验报告：

R 语言的高级数据管理

（2024-2025 学年第一学期）

学　　院　经济学院

课程名称　R 语言编程基础与金融数据分析

班　　级　22 金融 03 班

学　　号　24210819

姓　　名　周生瑞

任课教师　王皓

撰写日期　2024 年 10 月 13 日

代码地址：https://sr6688.github.io/

目录

图表索引

实验内容

- 实验目的
  - 了解 R 中常用的数据管理方法，熟悉基本的操作过程
- 实验内容
  - 对数据及进行变量重命名、缺失值分析、数据排序、随机抽样、变量计算等基本操作，熟悉 apply 函数族的使用
  - 编写一个函函数 stat，对数据进行描述性统计分析。
- 实验方法与步骤

1. 实验一：对于表格的数据集 CO2（在 R 中输入 CO2 即可查看）
   a) 查看数据集 CO2 中的变量名称，并将变量 Treatmen 的名称更改为 Treat
   b) 检验 CO2 中是否存在缺失值，检测缺失值的位置并删除含有缺失值的行。
   c) 对变量 uptake 按从大到小和从小到大排序，并对数据集 CO2 按照 uptake 排序（从大到小和从小到大）。
   d) 将 CO2 随机分成两组数据，第一组和第二组的比例为 6:4。
   e) 应用 tapply()函数，计算不同植物（Plant）对应的 uptake 的平均值
   f) 应用 aggregate 函数，计算不同植物（Plant）、不同类型（Type）对应的 uptake 的平均值
   g) 应用 lapply()函数，同时计算 conc 和 uptake 的均值
   h) 使用 grep()函数，查找出植物名称（Plant）中含有“Qn”的行的位置，并将这些行储存于变量 PlantQn 中
   i) 使用 gsub()函数，将 CO2 中植物名称（Plant）中的字符串“Qn”改为“QN”
2. 实验二：R 默认不提供函数计算峰度和偏度，可以自编公式或者使用 fBasics 包。加载 fasics 包，可使用 skewness（x）计算 x 的偏度，可使用 kurtosis（x）计算 x 的峰度。
   a) 编写函函数 stat，要求该函数同时计算均值、最大值、最小值、标准差、峰度和偏度。
   b) 生成自由度为 2 的 t 分布的 100 个随机数 t 并通过函数 stat 计算 t 的均值最大值，最小值，标准差、峰度和偏度

- 思考与实验总结
  - 对于一个新的位置的数据集，可以从哪些方面实现对数据的探索？
  - 如何通过数据管理得到实际情况中需要的数据集格式？

实验一

- 查看数据集 CO2 中的变量名称，并将变量 Treatmen 的名称更改为 Treat

data(CO2)

查看数据集 CO2 中的变量名称，并将变量 Treatment 的名称更改为 Treat。

names(CO2)

names(CO2)[names(CO2) == "Treatment"] <- "Treat"

```
> CO2
   Plant        Type      Treat conc uptake
1    Qn1      Quebec nonchilled   95   16.0
2    Qn1      Quebec nonchilled  175   30.4
3    Qn1      Quebec nonchilled  250   34.8
4    Qn1      Quebec nonchilled  350   37.2
5    Qn1      Quebec nonchilled  500   35.3
6    Qn1      Quebec nonchilled  675   39.2
7    Qn1      Quebec nonchilled 1000   39.7
8    Qn2      Quebec nonchilled   95   13.6
9    Qn2      Quebec nonchilled  175   27.3
10   Qn2      Quebec nonchilled  250   37.1
```

图 1 数据集 CO2

- 检验 CO2 中是否存在缺失值，检测缺失值的位置并删除含有缺失值的行。

which(is.na(CO2))

CO2 <- na.omit(CO2)

- 对变量 uptake 按从大到小和从小到大排序，并对数据集 CO2 按照 uptake 排序（从大到小和从小到大）。

CO2_desc <- CO2[order(-CO2$uptake),]

```
> CO2_desc
   Plant        Type      Treat conc uptake
21   Qn3      Quebec nonchilled 1000   45.5
14   Qn2      Quebec nonchilled 1000   44.3
20   Qn3      Quebec nonchilled  675   43.9
19   Qn3      Quebec nonchilled  500   42.9
35   Qc2      Quebec    chilled 1000   42.4
18   Qn3      Quebec nonchilled  350   42.1
11   Qn2      Quebec nonchilled  350   41.8
13   Qn2      Quebec nonchilled  675   41.4
42   Qc3      Quebec    chilled 1000   41.4
12   Qn2      Quebec nonchilled  500   40.6
17   Qn3      Quebec nonchilled  250   40.3
```

图 2 CO2 按照 uptake 排序（从大到小）

CO2_asc <- CO2[order(CO2$uptake),]

```
> CO2_asc
   Plant         Type      Treat conc uptake
71   Mc2 Mississippi    chilled   95    7.7
29   Qc2       Quebec    chilled   95    9.3
64   Mc1 Mississippi    chilled   95   10.5
43   Mn1 Mississippi nonchilled   95   10.6
78   Mc3 Mississippi    chilled   95   10.6
57   Mn3 Mississippi nonchilled   95   11.3
72   Mc2 Mississippi    chilled  175   11.4
50   Mn2 Mississippi nonchilled   95   12.0
73   Mc2 Mississippi    chilled  250   12.3
75   Mc2 Mississippi    chilled  500   12.5
74   Mc2 Mississippi    chilled  350   13.0
```

图 3 CO2 按照 uptake 排序（从小到大）

- 将 CO2 随机分成两组数据，第一组和第二组的比例为 6:4。

```
set.seed(123) #  设置随机种子以保证结果可重复
index <- sample(1:nrow(CO2), size = 0.6 * nrow(CO2))
CO2_group1 <- CO2[index, ]
```

```
> CO2_group1
   Plant         Type      Treat conc uptake
31   Qc2       Quebec    chilled  250   35.0
79   Mc3 Mississippi    chilled  175   18.0
51   Mn2 Mississippi nonchilled  175   22.0
14   Qn2       Quebec nonchilled 1000   44.3
67   Mc1 Mississippi    chilled  350   18.9
42   Qc3       Quebec    chilled 1000   41.4
50   Mn2 Mississippi nonchilled   95   12.0
43   Mn1 Mississippi nonchilled   95   10.6
81   Mc3 Mississippi    chilled  350   17.9
25   Qc1       Quebec    chilled  350   34.6
69   Mc1 Mississippi    chilled  675   22.2
57   Mn3 Mississippi nonchilled   95   11.3
```

图 4 CO2（Group1）

CO2_group2 <- CO2[-index,]

```
> CO2_group2
   Plant         Type      Treat conc uptake
1    Qn1       Quebec nonchilled   95   16.0
2    Qn1       Quebec nonchilled  175   30.4
3    Qn1       Quebec nonchilled  250   34.8
4    Qn1       Quebec nonchilled  350   37.2
6    Qn1       Quebec nonchilled  675   39.2
11   Qn2       Quebec nonchilled  350   41.8
16   Qn3       Quebec nonchilled  175   32.4
17   Qn3       Quebec nonchilled  250   40.3
19   Qn3       Quebec nonchilled  500   42.9
20   Qn3       Quebec nonchilled  675   43.9
22   Qc1       Quebec    chilled   95   14.2
24   Qc1       Quebec    chilled  250   30.3
```

图 5 CO2（Group2）

- 应用 tapply()函数，计算不同植物（Plant）对应的 uptake 的平均值

tapply(CO2$uptake, CO2$Plant, mean)

```
> tapply(CO2$uptake, CO2$Plant, mean)
     Qn1      Qn2      Qn3      Qc1      Qc3      Qc2      Mn3      Mn2      Mn1
33.22857 35.15714 37.61429 29.97143 32.58571 32.70000 24.11429 27.34286 26.40000
     Mc2      Mc3      Mc1
12.14286 17.30000 18.00000
```

图 6 uptake 的平均值（tapply()）

- 应用 aggregate 函数，计算不同植物（Plant）、不同类型（Type）对应的 uptake 的平均值

aggregate(uptake ~ Plant + Type, data = CO2, FUN = mean)

```
> aggregate(uptake ~ Plant + Type, data = CO2, FUN = mean)
   Plant        Type   uptake
1    Qn1      Quebec 33.22857
2    Qn2      Quebec 35.15714
3    Qn3      Quebec 37.61429
4    Qc1      Quebec 29.97143
5    Qc3      Quebec 32.58571
6    Qc2      Quebec 32.70000
7    Mn3 Mississippi 24.11429
8    Mn2 Mississippi 27.34286
9    Mn1 Mississippi 26.40000
10   Mc2 Mississippi 12.14286
11   Mc3 Mississippi 17.30000
12   Mc1 Mississippi 18.00000
```

图 7 uptake 的平均值（aggregate ()）

- 应用 lapply()函数，同时计算 conc 和 uptake 的均值

lapply(CO2[c("conc", "uptake")], mean)

```
$conc
[1] 435

$uptake
[1] 27.2131
```

图 8 uptake 的平均值（lapply()）

- 使用 grep()函数，查找出植物名称（Plant）中含有“Qn”的行的位置，并将这些行储存于变量 PlantQn 中

Plant_Qn <- CO2[grep("Qn", CO2$Plant),]

```
> Plant_Qn
   Plant   Type      Treat conc uptake
1    Qn1 Quebec nonchilled   95   16.0
2    Qn1 Quebec nonchilled  175   30.4
3    Qn1 Quebec nonchilled  250   34.8
4    Qn1 Quebec nonchilled  350   37.2
5    Qn1 Quebec nonchilled  500   35.3
6    Qn1 Quebec nonchilled  675   39.2
7    Qn1 Quebec nonchilled 1000   39.7
8    Qn2 Quebec nonchilled   95   13.6
9    Qn2 Quebec nonchilled  175   27.3
10   Qn2 Quebec nonchilled  250   37.1
```

图 9 使用 grep()函数，查找出植物名称（Plant）中含有“Qn”的行的位置

- 使用 gsub()函数，将 CO2 中植物名称（Plant）中的字符串“Qn”改为“QN”

CO2$Plant <- gsub("Qn", "QN", CO2$Plant)

```
> CO2$Plant <- gsub("Qn", "QN", CO2$Plant)
> CO2$Plant
 [1] "QN1" "QN1" "QN1" "QN1" "QN1" "QN1" "QN1" "QN2" "QN2" "QN2" "QN2" "QN2" "QN2" "QN2"
[15] "QN3" "QN3" "QN3" "QN3" "QN3" "QN3" "QN3" "Qc1" "Qc1" "Qc1" "Qc1" "Qc1" "Qc1" "Qc1"
[29] "Qc2" "Qc2" "Qc2" "Qc2" "Qc2" "Qc2" "Qc2" "Qc3" "Qc3" "Qc3" "Qc3" "Qc3" "Qc3" "Qc3"
[43] "Mn1" "Mn1" "Mn1" "Mn1" "Mn1" "Mn1" "Mn1" "Mn2" "Mn2" "Mn2" "Mn2" "Mn2" "Mn2" "Mn2"
[57] "Mn3" "Mn3" "Mn3" "Mn3" "Mn3" "Mn3" "Mn3" "Mc1" "Mc1" "Mc1" "Mc1" "Mc1" "Mc1" "Mc1"
[71] "Mc2" "Mc2" "Mc2" "Mc2" "Mc2" "Mc2" "Mc2" "Mc3" "Mc3" "Mc3" "Mc3" "Mc3" "Mc3" "Mc3"
```

图 10 使用 gsub()函数，将植物名称（Plant）中的字符串“Qn”改为“QN”

实验二

- 编写函函数 stat，要求该函数同时计算均值、最大值、最小值、标准差、峰度和偏度。
- 生成自由度为 2 的 t 分布的 100 个随机数 t 并通过函数 stat 计算 t 的均值最大值，最小值，标准差、峰度和偏度

```
library(fBasics)
stat <- function(x) {
  list(
    mean = mean(x),
    max = max(x),
    min = min(x),
    sd = sd(x),
    skewness = skewness(x),
    kurtosis = kurtosis(x)
  )
}
t_data <- rt(100, df = 2)
t_stats <- stat(t_data)
print(t_stats)
```

```
$mean
[1] -0.347204

$max
[1] 6.934788

$min
[1] -30.38759

$sd
[1] 3.553737

$skewness
[1] -5.851303
attr(,"method")
[1] "moment"

$kurtosis
[1] 48.55671
attr(,"method")
[1] "excess"
```

实验思考与经验总结

实验思考

- 对于一个新的位置的数据集，可以从哪些方面实现对数据的探索？

对于一个新的位置的数据集，数据探索（Exploratory Data Analysis，EDA）是一个至关重要的步骤，可以帮助我们理解数据集的特征、分布、异常值等。以下是一些常见的数据探索方面：

1. 数据清洗
   a) 检查缺失值并决定如何处理（删除、填充、插值等）。
   b) 识别并处理异常值或离群点。
   c) 确认没有错误的数据输入或记录。
2. 数据类型检查
   a) 确认每个变量的数据类型（数值型、分类型、日期时间型等）是否正确。
3. 基本描述统计
   a) 计算每个数值变量的均值、中位数、众数、最大值、最小值、范围、四分位数等。
   b) 计算分类变量的频数和百分比。
4. 分布分析
   a) 绘制直方图来观察数值变量的分布情况。
   b) 使用箱线图来识别数值变量的离群点。
   c) 检查分类变量的分布是否均衡。
5. 相关性分析
   a) 计算数值变量之间的相关系数（如皮尔逊、斯皮尔曼）。
   b) 绘制散点图矩阵来可视化变量之间的关系。
   c) 使用热图来展示变量间的相关性。
6. 多变量分析
   a) 使用主成分分析（PCA）或因子分析来降低数据的维度。
   b) 探索不同变量组合的影响。

7. 时间序列分析（如果数据集包含时间戳）
    a) 分析数据随时间的变化趋势。
    b) 识别季节性模式或周期性变化。
8. 空间数据分析（如果数据集包含地理位置信息）
    a) 使用地图可视化数据的地理分布。
    b) 分析地理位置与数据变量之间的关系。
9. 数据可视化
    a) 利用条形图、饼图、折线图等来展示数据的分布和趋势。
    b) 创造交互式可视化，如使用 Shiny 或 Tableau。
10. 模型诊断
    a) 如果已经建立了预测模型，检查模型的假设是否得到满足。
    b) 分析模型的残差，检查是否有模式或异常。
11. 数据集的完整性和一致性
    a) 确保数据集在逻辑上是完整的，没有遗漏或不一致的信息。
12. 数据集的可解释性
    a) 理解数据集中每个变量的含义和它们如何与研究问题相关联。

在进行数据探索时，通常需要使用统计软件和编程语言（如 R、Python 等），以及各种可视化工具。数据探索是一个迭代的过程，可能需要多次调整和重新分析以获得深入的洞察。

经验总结

- 如何通过数据管理得到实际情况中需要的数据集格式？

在实际情况中，数据集往往需要经过一系列的数据管理步骤才能转换成适合分析的格式。以下是一些关键的数据管理步骤，可以帮助你将原始数据转换成适合分析的数据集格式：

1. 数据导入
    a) 将数据从各种来源（如 CSV、Excel、数据库、API 等）导入到数据处理软件中。
2. 数据清洗

a) 处理缺失值决定是删除含有缺失值的行/列，还是通过插值、预测模型等方法填充缺失值。

b) 识别和处理异常值这些可能是由于错误输入、测量误差或其他非典型情况造成的。

3. 数据转换

a) 数据类型转换确保所有列的数据类型正确（例如，将日期字符串转换为日期类型）。

b) 规范化或标准化数值数据对数据进行缩放，使其位于相同的尺度上，便于比较和分析。

c) 分类变量编码将文本标签转换为数值编码，以便进行数学运算。

4. 数据整合

a) 合并数据如果数据分散在多个文件或表格中，需要将它们合并成一个统一的数据集。

b) 关联数据使用外键将不同数据源的数据关联起来。

5. 数据聚合

a) 对数据进行分组和聚合，以计算总和、平均值、最大值等统计量。

6. 特征工程

a) 创建新的特征基于现有数据生成新的变量，这些变量可能对分析更有用。

b) 特征选择选择最相关的特征进行分析，去除不必要的特征以减少噪声。

7. 数据划分

a) 将数据集划分为训练集和测试集，特别是在机器学习项目中。

8. 数据排序

a) 根据一个或多个变量对数据进行排序，以便于分析和理解数据的分布。

9. 数据抽样

a) 如果数据集非常大，可能需要进行随机抽样以减少数据量，同时保持数据的代表性。

10. 数据文档化

a) 记录数据的来源、处理步骤、使用的算法等，以便于复查和复现结果。

11. 数据质量检查

a) 确保数据的准确性、一致性和完整性。

12. 数据备份

a) 在进行任何重大更改之前，备份原始数据和中间步骤的结果。

13. 使用适当的工具和语言

a) 根据数据的复杂性和分析的需求，选择合适的数据处理工具（如 Excel、SQL、Python 的 Pandas 库、R 语言等）。

通过这些步骤可以将原始数据转换成一个干净、结构化、适合分析的数据集。这不仅有助于提高分析准确性，还可以节省后续分析的时间和努力。数据管理是一个迭代的过程，可能需要多次调整和优化。