第 6 章

数据清洗

在海量数据中不可避免地存在“脏”的数据，如何将这些“脏”数据有效转化成高质量的干净数据，涉及数据清理。数据清洗的最终目的就是提高数据的质量。数据的质量体现出数据的价值，更是知识服务水平的保障。“脏”数据主要包括：缺失数据、异常数据、重复数据、不一致的数据、敏感数据、虚假数据等，使分析结果更客观、更可靠。

本章实验使用的数据集是 titanic_train.csv。

6.1 缺失值分析

6.1.1 缺失值检测

（1）判断 x 是否为缺失值的函数是 is.na(x)，如果是则返回 TRUE，否则返回 FALSE。

【例 6.1】对 titanic_train.csv 缺失值分析。

```
>df<-read.table(file = "titanic_train.csv",header = TRUE,sep = ",",na.strings = "")
#读数据
head(df)                              #查看数据
PassengerId Survived Pclass                                                 Name
1           1        0      3                                Braund, Mr. Owen Harris
2           2        1      1      Cumings, Mrs. John Bradley (Florence Thayer)
3           3        1      3                                Heikkinen, Miss. Laina
4           4        1      1          Futrelle, Mrs. Jacques Heath (Lily May Peel)
5           5        0      3                              Allen, Mr. William Henry
6           6        0      3                                      Moran, Mr. James
```

```
  Sex    Age SibSp Parch      Ticket    Fare Cabin Embarked
1 male   22  1     0      A/5 21171  7.2500 <NA>  S
2 female 38  1     0      PC 17599  71.2833 C85   C
3 female 26  0     0 STON/O2. 3101282  7.9250 <NA>  S
4 female 35  1     0      113803    53.1000 C123  S
5 male   35  0     0      373450    8.0500 <NA>  S
6 male   NA  0     0      330877    8.4583 <NA>  Q
```

（2）缺失值分布可视化，如图 6.1 所示。

```
>library(mice)
>md.pattern(df)    #方法 1
```

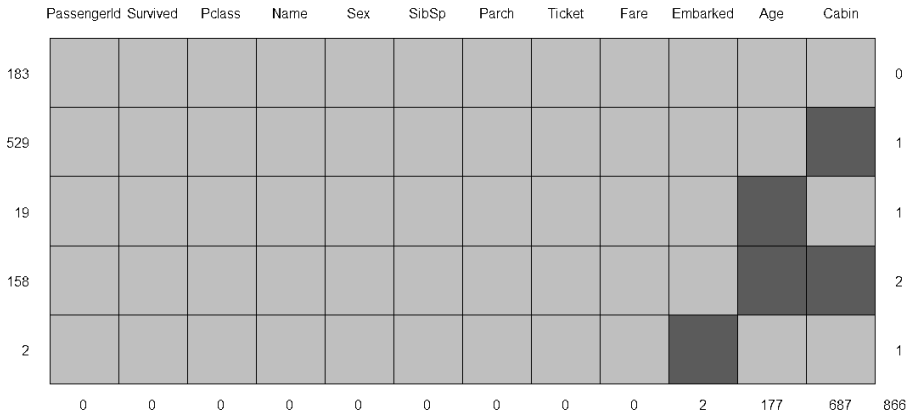

图 6.1　缺失值分布可视化

从图 6.1 可知：Embarked、Age 和 Cabin 3 个变量有缺失值，缺失值个数分别为 2、177、687。没有缺失值的样本是 183 个，只有 Cabin 有缺失值的样本数为 529 个，只有 Age 有缺失值的样本数为 19 个，只有 Embarked 有缺失值的样本数为 2 个，Age 和 Cabin 同时有缺失值的样本数为 158 个。

（3）统计每个变量的缺失值数量。

```
>summary(df)
```

（4）统计特定变量缺失值总数。

```
>sum(is.na(df$Age) == TRUE)
 [1] 177
```

6.1.2　缺失数据处理

对于缺失数据处理通常有以下 3 种方法。

方法 1：当缺失数据较少时直接删除相应样本。

删除缺失数据样本的前提是缺失数据的比例较少，而且缺失数据是随机出现的，这样删除缺失数据后对分析结果影响不大。

方法 2：对缺失数据进行插补。

用变量均值或中位数来代替缺失值，其优点在于不会减少样本信息，处理简单。其缺点在于当缺失数据不是随机出现时会产成偏误。

```
>df$Age<-ifelse(is.na(df$Age),mean(df$Age,na.rm =TRUE),df$Age) #空值替换
>sum(is.na(df$Age) == TRUE)
```

对出发港口缺失值处理：

```
>table(df$Embarked,useNA = "always")                    #出发港口的分布
>#将其中两个缺失值处理为概率最大的两个港口
>df$Embarked[which(is.na(df$Embarked))] = 'S'
>table(df$Embarked,useNA = "always")
```

6.2　异常值分析

异常值（离群点）是指测量数据中的随机错误或偏差，包括错误值或偏离均值的孤立点值。在数据处理中，异常值会极大地影响回归或分类的效果。

为了避免异常值造成的损失，需要在数据预处理阶段进行异常值检测。另外，某些情况下，异常值检测也可能是研究的目的，如数据造假的发现、电脑入侵检测等。

6.2.1　箱线图检测离群点

在一条数轴上，以数据的上下四分位数（Q1～Q3）为界画一个矩形盒子（中间 50%的数据落在盒内）；在数据的中位数位置画一条线段为中位线；默认延长线不超过盒长的 1.5 倍，延长线之外的点认为是异常值（用○标记），如图 6.2 所示。

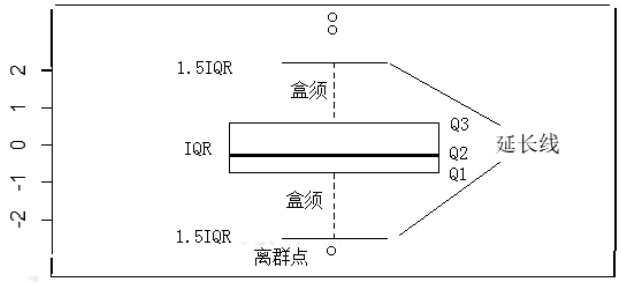

图 6.2　箱线图检测离群点

检测数据的异常值使用的函数是 boxplot.stats()，使用的数据集是 titanic_train.csv，执行如下代码得到的结果如图 6.3 所示。

```
y<-boxplot.stats(df[,'Fare'], coef=1.5, do.conf=TRUE, do.out=TRUE)
boxplot(df[,'Fare'])                                    #绘制箱线图
```

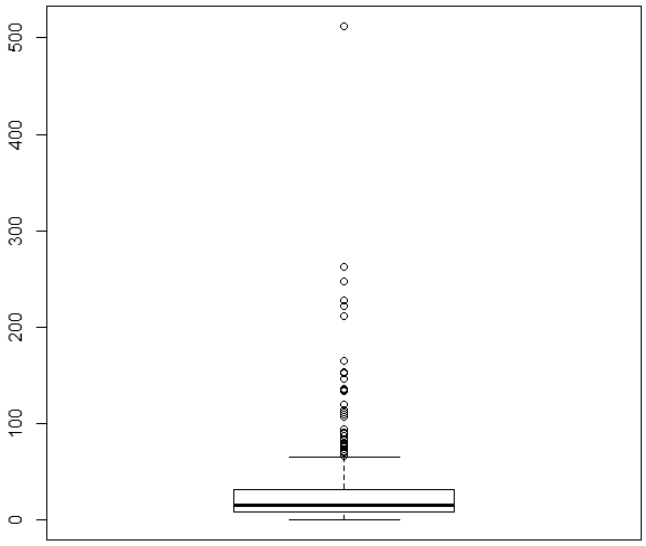

图 6.3　箱线图离群点检测

其中，coef 为盒须的长度是 IQR 的几倍，默认为 1.5 倍；do.conf 和 do.out 设置是否输出 conf 和 out。

返回值：stats 返回 5 个元素的向量值，包括盒须最小值、盒最小值、中位数、盒最大值、盒须最大值；n 返回非缺失值的个数；conf 返回中位数的 95%置信区间；out 返回异常值。

想查看具体的异常值，执行如下代码：

```
> y$out
```

想查看置信区间，执行如下代码：

```
>y$conf
[1] 13.23202 15.67638
```

6.2.2　点图检测离群点

点图通过离群点检测异常值。执行如下代码，结果如图 6.4 所示。

```
>set.seed(2016)
>x<-rnorm(100)
>y<-rnorm(100)
>df<-data.frame(x,y)
```

寻找异常值的 x 坐标。

```
a<-which(x %in% boxplot.stats(x)$out)
[1] 5
```

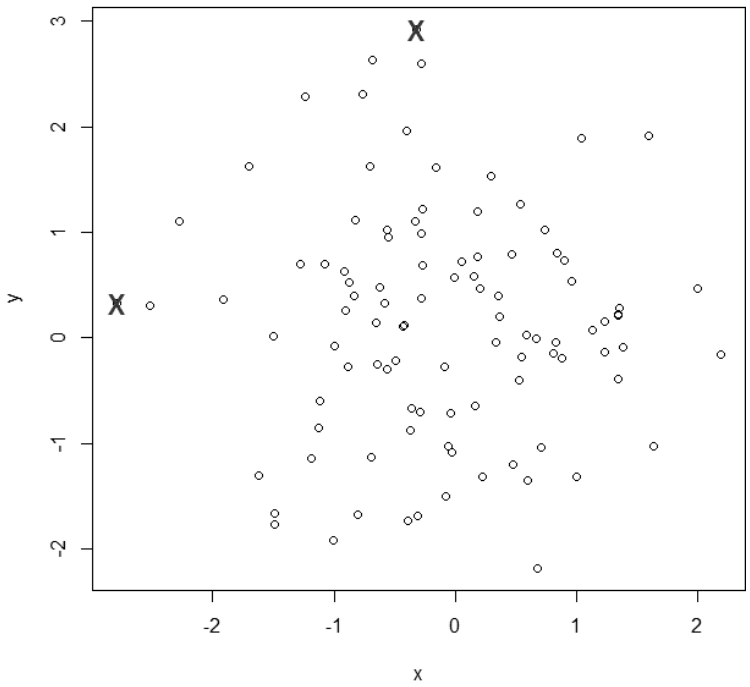

图 6.4　点图异常值检测

寻找异常值的 y 坐标。

```
>b<-which(y %in% boxplot.stats(y)$out)
>b
[1] 23
>plot(df)                                    #绘制 x、y 的点图
>p2<-union(a,b)                              #寻找变量 x 或 y 为异常值的坐标位置
>points(df[p2,],col="red",pch="x",cex=2)     #标记异常值
```

6.3　数据去重

数据重复检测函数包括 unique()和 duplicated()。

unique()函数只能作用于向量。

duplicated()函数是一个用来解决向量或者数据框重复值的函数，它会返回一个 TRUE 或 FALSE 的向量，以标注该索引所对应的值是否是前面数据所重复的值。

```
>test <- data.frame(
   x1 = c(1,2,3,4,5,1,3,5),
   x2 = c("a","b","c","d","e","a","b","e"),
   x3 = c("a","b","c","d","e","a","c","e"))
> test
```

```
  x1 x2 x3
1  1  a  a
2  2  b  b
3  3  c  c
4  4  d  d
5  5  e  e
6  1  a  a
7  3  b  c
8  5  e  e
> test[!duplicated(test),]                    #删掉所有列上都重复的
  x1 x2 x3
1  1  a  a
2  2  b  b
3  3  c  c
4  4  d  d
5  5  e  e
7  3  b  c
```

6.4 规范化

6.4.1 数据的中心化

数据的中心化是指数据集中的各项数据减去数据集的均值。

例如，有数据集 1，2，3，6，3，其均值为 3，那么中心化之后的数据集为−2，−1，0，3，0。

在 R 语言中可以使用 scale()方法来对数据进行中心化。

```
> data <- c(1, 2, 3, 6, 3)
> scale(data, center=T,scale=F)              #数据中心化
     [,1]
[1,]   -2
[2,]   -1
[3,]    0
[4,]    3
[5,]    0
attr(,"scaled:center")
[1] 3
```

6.4.2 数据标准化

数据标准化是指中心化之后的数据再除以数据集的标准差，即数据集中的各项数据减去数据集的均值再除以数据集的标准差。

例如，有数据集 1，2，3，6，3，其均值为 3，其标准差为 1.87，那么标准化之后的数据集为(1−3)/1.87，(2−3)/1.87，(3−3)/1.87，(6−3)/1.87，(3−3)/1.87，即−1.069，−0.535，0，1.604，0。

数据中心化和标准化的意义是一样的，目的都是为了消除量纲对数据结构的影响。R 语言使用 scale()方法来对数据进行标准化。

```
> data <- c(1, 2, 3, 6, 3)
> scale(data, center=T,scale=T)              #数据标准化
             [,1]
[1,] -1.06904
[2,] -0.53452
[3,]    0.00000
[4,]    1.60357
[5,]    0.00000
>attr(,"scaled:center")
[1] 3
>attr(,"scaled:scale")
[1] 1.8708
```

小数定标规范化：移动变量的小数点位置来将变量值映射到[−1,1]。

```
>#小数定标规范化
>i1=ceiling(log(max(abs(data[,1])),10))     #小数定标的指数
>c1=data[,1]/10^i1
>i2=ceiling(log(max(abs(data[,2])),10))
>c2=data[,2]/10^i2
>i3=ceiling(log(max(abs(data[,3])),10))
>c3=data[,3]/10^i3
>i4=ceiling(log(max(abs(data[,4])),10))
>c4=data[,4]/10^i4
>data_dot=cbind(c1,c2,c3,c4)
>#打印结果
>options(digits = 4)                        #控制输出结果的有效位数
>data_dot
```

6.5　格式转换

格式转换除类型转换，本节主要讲解的是字符串变换。其是数据清洗最难处理的内容，因为字符串如何变换与分析目标和业务场景有关，强依赖于人的经验。

如图 6.5 所示的 Titanic 数据集中 name 列信息可知，name 包含称谓、家族、名等信息。

```
> head(df)
  PassengerId Survived Pclass                                                 Name
1           1        0      3                              Braund, Mr. Owen Harris
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
3           3        1      3                               Heikkinen, Miss. Laina
4           4        1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
5           5        0      3                             Allen, Mr. William Henry
6           6        0      3                                     Moran, Mr. James
```

图 6.5　Titanic 数据 name 列信息

【例 6.2】统计 Titanic 数据集中 name 列各个称谓出现的次数。

（1）将 name 转换成 character。

```
>df$Name = as.character(df$Name)
```

（2）按空格键、Enter 键、换行等空白符，分割 name 为单词列表。

```
>strsplit(df$Name,"\\s+")
```

strsplit()把字符串按照某个规则进行拆分，返回列表。\\s 表示空格、回车、换行等空白符，+号表示一个或多个的意思。

（3）列表转换为向量。

```
>unlist(strsplit(df$Name,"\\s+"))
```

（4）统计单词出现的频次。

```
>table_words = table(unlist(strsplit(df$Name,"\\s+")))
```

（5）筛选称谓。

```
>title<-grep("\\.",names(table_words))
```

用（"\\."）作为一种正则表达式及筛选的条件。

（6）对称谓出现次数排序。

```
>sort(table_words[title],decreasing = TRUE)
Mr.      Miss.      Mrs.    Master.       Dr.       Rev.       Col.     Major.
517       182       125        40         7         6          2         2
Mlle.    Capt. Countess.    Don. Jonkheer.     L.     Lady.     Mme.      Ms.      Sir.
2         1         1         1         1         1         1         1         1
```

【例 6.3】根据各个称谓的平均年龄填充年龄的缺失值。

（1）将称谓和年龄组成一个新的数据框 tb。

```
>library(stringr)
>tb = cbind(df$Age,str_match(df$Name,"[a-zA-Z]+\\."))
>tb
```

tb 的左侧列出了年龄包括默认值，右侧列出了正则表达式。

（2）筛选年龄为空值对应的称谓。

```
>tb[is.na(tb[,1]),2]
```

（3）对筛选后的称谓进行统计。

```
>table(tb[is.na(tb[,1]),2])
Dr.     Master.   Miss.    Mr.     Mrs.
1       4         36       119     17
```

（4）统计每类人的年龄均值（不包含缺失值）。

grepl()检索目标行命令， Mr\\.的\\表示绝对匹配。

```
>mean.mr = mean(df$Age[grepl("Mr\\.",df$Name)&!is.na(df$Age)])
>mean.mrs = mean(df$Age[grepl("Mrs\\.",df$Name)&!is.na(df$Age)])
>mean.dr = mean(df$Age[grepl("Dr\\.",df$Name)&!is.na(df$Age)])
>mean.miss = mean(df$Age[grepl("Miss\\.",df$Name)&!is.na(df$Age)])
>mean.master = mean(df$Age[grepl("Master\\.",df$Name)&!is.na(df$Age)])
```

（5）填充年龄缺失值。

如果某个人群包含缺失值，插补的方式是将每一类人群年龄平均值插补到缺失值中。

```
>df$Age[grepl("Mr\\.",df$Name)&is.na(df$Age)] = mean.mr
>df$Age[grepl("Mrs\\.",df$Name)&is.na(df$Age)] = mean.mrs
>df$Age[grepl("Dr\\.",df$Name)&is.na(df$Age)] = mean.dr
>df$Age[grepl("Miss\\.",df$Name)&is.na(df$Age)] = mean.miss
>df$Age[grepl("Master\\.",df$Name)&is.na(df$Age)] = mean.master
```

习题

一、单选题

1．判断是否有缺失值的函数是_____。

A．is.na()　　B．complete.cases()

C．NA()　　D．NULL()

2．数据重复检测函数中_____函数是用来解决向量或者数据框重复值的，并且它会返回一个 TRUE 或 FALSE 的向量。

A．duplicated()　　B．unique()

C．matrix()　　D．data frame()

3．数据集 1，2，3，6，3 经过中心化的结果是_____。

A．−2，−1，0，3，0　　B．−1，0，1，4，1

C．−3，−2，−1，2，−1　　D．1，2，3，6，3

二、填空题

1．对于缺失数据通常有三种应付手段：_____、_____和_____。

2．在 R 中，用代码 NA 表示缺失数据。在向量及数据框中，在缺失数

据处应使用该代码作为占位符。在 R 中对含有缺失值的向量进行计算，会返回一个包装缺失值的向量作为结果，例如：

```
> u=(3,5,6,NA,12,14)
>u
```

执行结果是_____。

```
>2^u
```

执行结果是_____。

3．检测数据的异常值时使用函数_____。

4．在 R 语言中，通常使用_____来画直方图。

5．当对数据进行批量操作时，可以通过对函数返回值进行约束，根据是否提示错误判断、是否存在数据不一致问题，可以通过_____函数。

6．数据集 1，2，3，6，3 经过数据的标准化后的结果是_____。

7．如果 x 有缺失值，则函数 is.na(x)=_____。

8．_____是通过变量间关系来预测缺失数据。

9．_____是指测量数据中的随机错误或偏差，包括错误值或偏离均值的孤立点值。

三、判断题

1．数据清洗的最终目的是提高数据的质量。（　　）

2．用变量均值或中位数来代替缺失值是最好的方法。（　　）

3．若 T 是一个矩阵，则 unique(T)的功能是识别重复行。（　　）

4．数据的中心化是指数据集中的各项数据加上数据集的均值。（　　）

5．数据标准化是指中心化之后的数据再除以数据集的标准差。（　　）

四、多选题

1．数据清理的主要任务是通过识别_____来“清理脏数据”。

A．缺失值　　B．噪声数据

C．不一致数据　　D．重复数据

2．数据重复检测函数包括_____。

A．unique()　　B．duplicated()

C．only()　　D．repeat()

五、简答题

1．简述缺失数据的处理方法。

2．简述异常值分析的常用方法。