

ElegantBook：优美的 LaTeX 书籍模板

ElegantLaTeX 经典之作

作者：Ethan Deng & Liam Huang
组织：ElegantLaTeX Program
时间：April 9, 2022
版本：4.3
自定义：信息

ElegantLaTeX Program

不要以为抹消过去，重新来过，即可发生什么改变。——比企谷八幡

目录

第 1 章 R 语言数据清理 **1**

1.1 分类的文本数据 1

1.2 压缩种类数量 1

第 1 章 R 语言数据清理

1.1 分类的文本数据

衣服尺寸是因子类型 (factor) 的数据，分别为 S,M,L 和 XL，但是因为一些原因会出现不在这四类的情况，我们需要处理这种数据。首先我们有预想的因子类型集合，这里是 S,M,L 和 XL。然后检测数据的某一列在不在这里面。利用找出在在因子集合内的所有数据，或者利用找出不在因子集合内的所有数据，然后进行处理。下面数据（从旧金山国际机场（SFO）乘坐航班的乘客的调查答复。）来自链接。

```
sfo_survey <- readRDS("sfo_survey_ch2_1.rds")
# 预设的因子集合
dest_sizes <- data.frame("dest_size"=c("Medium", "Small", "Large","Hub"))

#检验dest_size出现次数
sfo_survey %>% count(dest_size)

# 找到出错的 dest_size行
sfo_survey %>%
# 链接dest_sizes数据框来找到出错的dest_size行
anti_join(dest_sizes, by = "dest_size") %>%
#选取一部分查看
select(id, airline, destination, dest_size)

# 移除
sfo_survey %>%
semi_join(dest_sizes, by = "dest_size") %>%
#计算一下每种dest_size出现次数
count(dest_size)
```

如果是确切的大小写问题或者空格问题，可使用相应的函数。下面是在里面使用。

```
sfo_survey <- sfo_survey %>%
# dest_size_trimmed
mutate(dest_size_trimmed = str_trim(dest_size),
# cleanliness_lower
cleanliness_lower = str_to_lower(cleanliness))
```

1.2 压缩种类数量

有时候还解决不了问题，需要把一些出错的种类更改过来，又或者压缩成我们指定的几种种类，需要 **forcats** 包中的来协助我们。

```
install.packages("forcats")
library(forcats)
```