第四次课堂作业

斯蓬 220810332

2024-11-19

目录

1 数据读取和处理问题

```
# 1. 读取本地数据 data.csv
data <- read.csv("data.csv", header = TRUE, stringsAsFactors = FALSE)

# 2. 查看数据前 10 行
print(head(data, 10))
```

```
##          createTime education  salary
## 1  2020/3/16 11:30      本科 20k-35k
## 2  2020/3/16 10:58      本科 20k-40k
## 3  2020/3/16 10:46      不限 20k-35k
## 4  2020/3/16 10:45      本科 13k-20k
## 5  2020/3/16 10:20      本科 10k-20k
## 6  2020/3/16 10:33      本科 10k-18k
## 7  2020/3/16 10:11      硕士 16k-30k
## 8  2020/3/16 09:49      本科 10k-15k
## 9  2020/3/16 09:25      不限   6k-8k
```

```
## 10 2020/3/16 09:35      本科 12k-20k
```

```r
# 3. 读取 salary 列
if ("salary" %in% names(data)) {
  library(stringr)  # 加载 stringr 包用于字符串处理

  # 提取 salary 列的最小值和最大值
  salary_split <- str_extract_all(data$salary, "\\d+")  # 提取数字部分
  min_salary <- sapply(salary_split, function(x) ifelse(length(x) > 0, as.numeric(x[1]), NA))  # 最小
  max_salary <- sapply(salary_split, function(x) ifelse(length(x) > 1, as.numeric(x[2]), NA))  # 最大

  # 将提取的最小值和最大值作为新列
  data$min_salary <- min_salary
  data$max_salary <- max_salary

  # 计算平均值并替换 salary 列
  avg_salary <- rowMeans(cbind(min_salary, max_salary), na.rm = TRUE)
  data$salary <- avg_salary  # 替换 salary 列为平均值

  # 查看数据后 10 行
  print(" 前后 10 行（salary 列已处理为平均值）: ")
  print(head(data, 10))
  print(tail(data, 10))
}
```

```
## [1] "前后 10 行（salary 列已处理为平均值）："
##         createTime education salary min_salary max_salary
## 1  2020/3/16 11:30      本科   27.5         20         35
## 2  2020/3/16 10:58      本科   30.0         20         40
## 3  2020/3/16 10:46      不限   27.5         20         35
## 4  2020/3/16 10:45      本科   16.5         13         20
## 5  2020/3/16 10:20      本科   15.0         10         20
## 6  2020/3/16 10:33      本科   14.0         10         18
## 7  2020/3/16 10:11      硕士   23.0         16         30
## 8  2020/3/16 09:49      本科   12.5         10         15
```

```
## 9  2020/3/16 09:25      不限    7.0          6          8
## 10 2020/3/16 09:35      本科   16.0         12         20
##          createTime education salary min_salary max_salary
## 126 2020/3/16 11:13      本科   12.5         10         15
## 127 2020/3/16 11:12      本科    4.0          3          5
## 128 2020/3/16 09:44      硕士   12.5         10         15
## 129 2020/3/16 10:57      本科   22.5         15         30
## 130 2020/3/16 09:46      本科   20.0         15         25
## 131 2020/3/16 11:36      本科   14.0         10         18
## 132 2020/3/16 09:54      硕士   37.5         25         50
## 133 2020/3/16 10:48      本科   30.0         20         40
## 134 2020/3/16 10:46      本科   19.0         15         23
## 135 2020/3/16 11:19      本科   30.0         20         40
```

```r
# 4. 根据学历分组，计算平均工资并打印出来
if ("education" %in% names(data)) {
  library(dplyr)

  avg_salary_by_education <- data %>%
    group_by(education) %>%
    summarise(avg_salary = mean(salary, na.rm = TRUE))

  print(avg_salary_by_education)
}
```

```
## # A tibble: 4 x 2
##   education avg_salary
##   <chr>          <dbl>
## 1 不限            19.6
## 2 大专            10
## 3 本科            19.4
## 4 硕士            20.6
```