第五組

黎宏濬、林孝儒、許震浩、張立宏

2024-10-16

資料處理過程

1. 資料庫位置設定 (setwd)

根據不同使用者，設定工作目錄 (setwd) 到適當的 Dropbox 目錄位置，這樣便能夠從指定路徑讀取數據文件。

```
library(readxl)
library(dplyr)
library(data.table)
library(ggplot2)
# library(psych)

user <- Sys.info()["user"]

if (user == "brianhjli"){
  setwd("/Users/brianhjli/Dropbox/113-1/Research Methodology/DSE-5")
} else if (user == "QQ"){
  setwd("C:/Users/QQ/Dropbox/DSE-5")
} else if (user == "hayashijikyou"){
  setwd("/Users/hayashijikyou/Library/CloudStorage/Dropbox/DSE-5")
} else if (user == ""){
  setwd("")
}
```

2. 讀取資料 (fread)

使用 fread 函數從指定路徑讀取不同的 CSV 資料檔案，包括銷售數據、家庭收支統計、物價統計及薪資收入資料。

```
sale_data <- fread("data/    .csv")
# income_data <- fread("data/    .csv", fileEncoding = "Big5")
house_income_data <- fread("data/    .csv")
price_data <- fread("data/   .csv")
salary_data <- fread("data/    .csv")
```

3. 銷售數據處理 (sale_data)

- **更改欄位名稱**：將第二行作為欄位名稱，並刪除前兩行不需要的數據。
- **年份和月份提取**：利用 gsub 函數去掉「年」和「月」字，並將年份和月份轉換為數值型。
- **缺失值填補**：利用 nafill 函數按順序填補年份和月份的缺失值。
- **數據格式轉換**：將第三到第十二欄的數據轉換為數值型，並將千分位符號去掉。
- **重新命名**：將產品名稱及其對應的數據欄位改為英文名稱。

```r
colnames(sale_data) <- as.character(sale_data[2,])
sale_data <- sale_data[-1,]
sale_data <- sale_data[-1,]
sale_data <- sale_data[,1:12]

setnames(sale_data, 1, "year")
setnames(sale_data, 2, "month")
sale_data[, year := as.numeric(gsub(" ", "", year))]
sale_data[, month := as.numeric(gsub(" ", "", month))]
sale_data[sale_data == "-"] <- NA
sale_data[sale_data == ""] <- NA
sale_data[, 1:2 := lapply(.SD, nafill, type = "locf"), .SDcols = 1:2]
sale_data[, 3:12] <- lapply(sale_data[, 3:12], function(x) as.numeric(gsub(",", "", x)))

setnames(sale_data,
         c("(0920010)  ( )", "(0920910)  ( )", "(0920930)  ( )",
           "(0920940)  ( )", "(0920950)  ( )", "(0920010) ",
           "(0920910)  ", "(0920930)  ", "(0920940)  ", "(0920950)  "),
         c("FruitVegetableJuice_volumn", "CarbonatedBeverage_volumn", "SportsDrink_volumn",
           "CoffeeDrink_volumn", "TeaDrink_volumn", "FruitVegetableJuice_value",
           "CarbonatedBeverage_value", "SportsDrink_value", "CoffeeDrink_value", "TeaDrink_value"))
```

4. 家庭收入數據處理 (house_income_data)

- **更改欄位名稱**：將第二行設定為欄位名稱，刪除前兩行不需要的數據。
- **欄位命名**：對欄位重新命名，如 “year” 、“DisposableIncome” 等。
- **年份轉換**：將民國年轉換為西元年，並將所有收入數據轉換為數值型格式。

```r
colnames(house_income_data) <- as.character(house_income_data[2,])
house_income_data <- house_income_data[-1,]
house_income_data <- house_income_data[-1,]

colnames(house_income_data) <- c("year", "DisposableIncome", "AvgHouseholdDisposableIncome",
                  "LowestQuintileIncome", "SecondLowestQuintileIncome",
                  "MiddleQuintileIncome", "SecondHighestQuintileIncome",
                  "HighestQuintileIncome")
house_income_data[, year := as.numeric(gsub(" ", "", year))]
house_income_data[, year := year+1911]
house_income_data[, ] <- lapply(house_income_data[,], function(x) as.numeric(gsub(",", "", x)))
```

5. 物價數據處理 (price_data)

- **年份和月份提取**：將 “統計期” 分割為年份和月份，並轉換為數值型。
- **重新命名欄位**：將 “總指數” 和 “15.非酒精性飲料及材料” 重新命名為 “TotalIndex” 和 “SoftDrink” 。
- **年份轉換**：將民國年轉換為西元年。

```r
price_data[, c("year", "month") := tstrsplit(  , " ")]
price_data[, year := as.numeric(year)]
price_data[, month := as.numeric(gsub(" ", "", month))]
setnames(price_data, c("  ", "15.     "), c("TotalIndex", "SoftDrink") )
price_data <- price_data[, c("year", "month", "TotalIndex", "SoftDrink")]
price_data[, year := year+1911]
```

6. 薪資數據處理 (`salary_data`)

- **年份和月份提取**：將 “統計期” 分割為年份和月份，並轉換為數值型。
- **重新命名欄位**：將 “總薪資” 重新命名為 “TotalSalary” 。
- **年份轉換**：將民國年轉換為西元年。

```
salary_data[, c("year", "month") := tstrsplit(  , "  ")]
salary_data[, year := as.numeric(year)]
salary_data[, month := as.numeric(gsub(" ", "", month))]
setnames(salary_data, c("  "), c("TotalSalary"))
salary_data <- salary_data[, c("year", "month", "TotalSalary")]
salary_data[, year := year+1911]
```

7. 合併資料表

- **Full Join**：將銷售數據、物價數據、薪資數據進行 `full join`，然後再將家庭收入數據根據 “year” 進行合併。

```
list_of_dts <- list(sale_data, price_data, salary_data)
full_join_all <- Reduce(function(x, y) merge(x, y, by = c("year", "month"), all = TRUE), list_of_dts)
full_join_all <- merge(full_join_all, house_income_data, by = "year", all = TRUE)

full_join_all[, FruitVegetableJuice_price := FruitVegetableJuice_value/FruitVegetableJuice_volumn]
full_join_all[, CarbonatedBeverage_price := CarbonatedBeverage_value/CarbonatedBeverage_volumn]
full_join_all[, SportsDrink_price := SportsDrink_value/SportsDrink_volumn]
full_join_all[, CoffeeDrink_price := CoffeeDrink_value/CoffeeDrink_volumn]
full_join_all[, TeaDrink_price := TeaDrink_value/TeaDrink_volumn]

print(full_join_all)
```

8. 結果輸出 (`write.csv`)

將合併後的完整數據集保存為 CSV 檔案。

```
write.csv(full_join_all, "data/fullData.csv", row.names = FALSE)
```

總結

這段代碼的核心是讀取多個不同來源的數據集，進

行必要的清理和格式化，然後合併為一個完整的數據集，並且最終保存為 CSV 檔案，方便後續的分析使用。