

Untitled

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。

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。

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（果蔬汁、碳酸飲料等）及其對應的數據欄位分別改為英文名稱，例如“FruitVegetableJuice_volumn”和“FruitVegetableJuice_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，方便後續操作。
- 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（去掉逗號）。

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。同樣使用 gsub 去掉「月」字。
- 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。

6. 薪資數據處理 (salary_data)

- 年份和月份提取：同樣使用 tstrsplit 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。

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，也就是透過相同的年份和月份進行合併。接著，再將家庭收入數據根據“year”進行合併，這是一個按列的全合併操作，確保所有資料保留，即使某些年份和月份不完全匹配。

8. 結果輸出 (write.csv)

- 最後，將合併後的完整數據集保存為 CSV 檔案，方便後續分析和使用。

總結

這段代碼的核心是讀取多個不同來源的數據集，進行資料清理（例如去掉符號、填補缺失值、年份轉換等），並通過 `merge` 函數將它們合併在一起。最終，將合併後的完整數據集寫入 CSV 檔案進行保存。