Problem Set 4

Chang Gao

1. Data Scraping

I am currently working with a panel dataset to study housing prices in China. This dataset primarily contains monthly real estate transaction data for various cities from 2007-2024, including the number of housing units sold, total floor area, and monthly average prices for each city. However, the monthly price data is incomplete, only available for some of the cities.

Recently, my friend informed me that China's National Bureau of Statistics publishes monthly relative price changes for 70 representative cities. An example (in Chinese) can be found at https://www.stats.gov.cn/sj/zxfb/202412/t20241216_1957755.html. I noticed that these tables are stored in HTML documents and are downloadable, which is convenient. These HTML files have been published monthly since 2015.

I have been downloading these HTML files and plan to use AI chatbots, particularly Claude.ai, to help write Python or R code for data extraction, translate the cities names into English. And merge it with my existing dataset, although that the two datasets cover different cities.

In my experience, Claude.ai 3.5's coding capability is better than DeepSeek, ChatGPT's free version, and Google Gemini's trial paid version. While DeepSeek is better than other chatbots in writing text in Chinese.

2. Q6

(7)

```
> class(df1)
[1] "tbl_df"     "tbl"        "data.frame"

> class(df)
[1] "tbl_spark"  "tbl_sql"    "tbl_lazy"   "tbl"
```

(8) They are the same?

```
> names(df1)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> colnames(df)
[1] "Sepal_Length" "Sepal_Width"  "Petal_Length" "Petal_Width"  "Species"
```

(13b) Here is what I get

```
> df2 %>% arrange(Species) %>% head %>% print
# Source:     SQL [?? x 3]
# Database:   spark_connection
# Ordered by: Species
Species     mean count
<chr>      <dbl> <dbl>
1 setosa      5.01    50
2 versicolor  5.94    50
3 virginica   6.59    50
```