

README

Dongwook Kim

2025-07-29

Contents

1 KEAPS Self-employment Rate

This project analyzes **self-employment rates** using Korean Economically Active Population Survey (KEAPS) from 1981 to 2024. It calculates and visualizes trends by **year, industry(KSIC), age, and gender**.

1.1 Overview

- **Data source:** KEAPS (Korean Economically Active Population Survey)
 - Available at [MDIS] (<https://mdis.kostat.go.kr/eng/index.do>)
- **Scope**
 - **Period:** 1981 - 2024 (monthly KEAPS microdata)
 - **Unit of analysis:** Individuals of age 15+
 - **Key variables:**
 - * Employment status, industry (KSIC), gender & age group
 - * Self-employment classification
 - **Type 1 (OECD):** Family workers + self-employed with and without employees
 - **Type 2 (KOSIS):** Self-employed with and without employees
 - **Type 3 (Project):** Self-employed without employees only
 - **Aggregation level:**
 - * Yearly trends
 - * Industry-level trends
 - * Age & Gender group trends
- **Outputs**
 - **Monthly Outputs**
 - * `month_trends_age_group_2.rdata`, etc.: monthly trends
 - `month_overall`: monthly self-employment rates
 - `month_ind`: monthly rates by industry
 - `month_as`: monthly rates by age & gender group
 - `month_as_wg`: monthly rates within age & gender group
 - `month_supp`: monthly labor participation & unemployment rates
 - **Yearly Outputs**
 - * `year_trends_age_group_2.rdata`, etc.: yearly trends

- `year_overall`: yearly self-employment rates
- `year_ind`: yearly rates by industry
- `year_as`: yearly rates by age & gender group
- `year_as_wg`: yearly rates within age & gender group
- `year_supp`: yearly labor participation & unemployment rates

1.2 Project Structure

Project/

```
R/
  raw/                      # Raw .R scripts from KOSIS
  modules/
    preprocess/             # Preprocessing functions
    cleaning/               # Variable cleaning & standardization
    calc/                   # Calculation of self-employment rates
    visualization/          # Plotting functions

  main_pipeline.R
  01_preprocess.R
  02_data_clean.R
  03_rate_calculation.R
  03_data_visualization.R # Will be uploaded soon

data/
  raw/                      # Raw KEAPS .csv files
  processed/                # Processed .rds files
  cleaned/                  # Cleaned .rds files
  output/                   # Final monthly/yearly .rdata files

image/
  README.md                 # This documentation
```

1.3 How to Run

1. Place raw .csv files in `data/raw/`
2. Place KOSIS .R scripts in `R/raw/`
3. In RStudio (project opened with `.Rproj`), open `main_pipeline.R`
4. Run all sections
5. Outputs will be saved in `data/output/` as .rdata files
6. Images will be saved in `image/`

1.4 Analysis Pipeline

1. **Process raw data**
 - Raw .csv files are processed with KOSIS-provided scripts
2. **Load raw data**
 - .rds files from 1981–2024 stored in `data/processed/`
3. **Standardize variables**

- Aligns consistent variable names across years

4. Recode variables

- Translates Korean -> English
- Groups ages hierarchically:
 - 5-year intervals
 - 60+, 70+, 80+ categories
- Classifies self-employment into 3 types (see above)

5. Calculate monthly rates

- Self-employment rate (overall, by industry, by/within group)
- Supplementary: labor force participation & unemployment rate

6. Aggregate to yearly outputs

- Averages monthly results by year

1.5 Dependencies

- **R** (>= 4.4.1)
- **Required Packages:**
 - dplyr
 - stringr
 - rlang
 - purrr
 - stringi
 - data.table
 - readr
 - ggplot2
- **Optional:** grafify for color palettes
- To install the packages, use in R:

```
install.packages(c("dplyr", "stringr", "rlang", "purrr", "stringi", "data.table", "readr", "ggplot2"))
```

1.6 Example Output

Below is an example plot showing yearly trends of self-employment rate (1981-2024). Each colored line shows different types of self-employment as defined above. The dashed line represents tax-data comparable estimates which excludes family workers from the employed population. A vertical red line indicates 1997 IMF Fi-