PROGRAMMING

Jana Flakova

2025-06-18

Part 2 - Data Sourcing

2.1 Load in the Data

We'll be using two datasets in this project:

1. A custom dataset on the Gender Pay Gap in the EU (loaded from a CSV)
2. A dataset on sectoral wages and gender workforce distribution in Germany (2023)

```
library(tidyverse)
library(dplyr)
library(ggplot2)
```

```
TableCountries = read.csv("data/TableCountries.csv")
Germany_subgroup = read.csv("data/Germany_subgroup/Germany_subgroup.csv")
```

2.2 Short Summary

```
head(TableCountries)
```

```
##   X country year gender_pay_gap monthly_income  p_female    p_male
## 1 1 Austria 2021           19.1       4013.833 0.4097419 0.5902581
## 2 2 Austria 2022           18.7       4228.417 0.4135149 0.5864851
## 3 3 Austria 2023           18.3       4542.333 0.4117647 0.5882353
## 4 4 Belgium 2021            1.8       4141.333 0.4285521 0.5714479
## 5 5 Belgium 2022            0.7       4451.500 0.4338630 0.5661370
## 6 6 Belgium 2023            0.7       4832.417 0.4315789 0.5684211
```

```
head(Germany_subgroup)
```

```
##                                           sector income_eur gpg_2023
## 1               Financial and insurance activities       5841       26
## 2                    Information and communication       5769       21
## 3 Professional, scientific and technical activities       5436       26
## 4  Electricity, gas, steam, air conditioning supply       5352       14
## 5                                      Education       4733        9
## 6                           Mining and quarrying       4544        2
```

```
##                         occupation_category p_female p_male
## 1                             Professionals    0.501  0.499
## 2                             Professionals    0.501  0.499
## 3                             Professionals    0.501  0.499
## 4 Technicians and associate professionals    0.537  0.463
## 5                             Professionals    0.501  0.499
## 6        Craft and related trades workers    0.102  0.898
```

2.3 Metadata Description

Main Dataset: EU Gender Pay Gap Data (2021–2023)

- **Title**: EU Gender Pay Gap and Workforce Composition
- **Years Covered**: 2021, 2022, 2023
- **Geographic Scope**: All EU countries + EU-27 aggregate
- **Sources**: Eurostat, EIGE
- **Units**:
  - `monthly_income`: Euros (gross, monthly)
  - `gender_pay_gap`: Percentage (%)
  - `p_female`, `p_male`: Share in workforce (0–1)
- **Variables**:
  - `country`: Country name (categorical)
  - `year`: Observation year (numeric)
  - `gender_pay_gap`: Unadjusted gender pay gap
  - `monthly_income`: Average monthly gross income
  - `p_female` / `p_male`: Workforce gender shares
- **Collection Method**: Administrative statistics, not survey-based

---

Subgroup Dataset: Germany by Sector (2023)

- **Title**: Sectoral Gender Pay Gap and Earnings – Germany
- **Year**: 2023 (with income data likely from 2022)
- **Scope**: Sector-level data for Germany
- **Sources**: Statista, Destatis

- **Units**:
  - `income_eur`: Euros (gross, monthly)
  - `gpg_2023`: Percentage (%)
  - `p_female`, `p_male`: Share in workforce (0–1)
- **Variables**:
  - `sector`: Economic sector name (categorical)
  - `income_eur`: Average monthly income
  - `gpg_2023`: Gender pay gap
  - `occupation_category`: Dominant occupational group
  - `p_female` / `p_male`: Workforce gender shares
- **Collection Method**: Aggregated administrative data