Gender Pay Inequality in the EU: Sectoral and Regional Analysis

Jana Fľaková 2856763,

2025-06-20

Names: Jana Fľaková **Tutorial Group**: add pls **Lecturer**: add pls

Part 1 - Identify a Social Problem

1.1 Describe the Social Problem

Aliya

References:
ALIYA

Part 2 - Data Sourcing

2.1 Load in the Data

```
# Read CSVs from project GitHub repo structure
TableCountries = read_csv("data/TableCountries.csv")
Germany_subgroup = read_csv("data/Germany_subgroup/Germany_subgroup.csv")
```

2.2 Summary of the Dataset

```
head(TableCountries)
```

```
## # A tibble: 6 x 7
##    ...1 country  year gender_pay_gap monthly_income p_female p_male
##   <dbl> <chr>   <dbl>          <dbl>          <dbl>    <dbl>  <dbl>
## 1     1 Austria  2021           19.1          4014.    0.410  0.590
## 2     2 Austria  2022           18.7          4228.    0.414  0.586
## 3     3 Austria  2023           18.3          4542.    0.412  0.588
## 4     4 Belgium  2021            1.8          4141.    0.429  0.571
## 5     5 Belgium  2022            0.7          4452.    0.434  0.566
## 6     6 Belgium  2023            0.7          4832.    0.432  0.568
```

The EU-wide data set includes country-level variables such as the gender pay gap, average monthly income, and male and female share in workforce for 2021–2023. The Germany subgroup includes sector-level earnings, gender pay gaps, and gender share in workforce.

2.3 Describe the Type of Variables

The datasets are compiled from administrative sources such as Eurostat and Destatis. Variables include:

- `gender_pay_gap`: Percentage difference in male vs female income
- `monthly_income`: Average monthly gross pay (€)
- `p_female` / `p_male`: Employment shares by gender
- `income_eur`: Average monthly income by sector in Germany
- `gpg_2023`: Sector-specific gender pay gap

Part 3 - Quantifying

3.1 Data Cleaning

```
TableCountries$year = as.factor(TableCountries$year)
Germany_subgroup$income_quartile = cut(Germany_subgroup$income_eur,
  breaks = quantile(Germany_subgroup$income_eur, probs = seq(0, 1, 0.25), na.rm = TRUE),
  include.lowest = TRUE,
  labels = c("Q1 (lowest)", "Q2", "Q3", "Q4 (highest)"))
```

3.2 Generate Necessary Variables

```
# EU country-level income estimation
male_income = numeric(nrow(TableCountries))
female_income = numeric(nrow(TableCountries))
for (i in 1:nrow(TableCountries)) {
  income = TableCountries$monthly_income[i]
  gpg = TableCountries$gender_pay_gap[i] / 100
  p_f = TableCountries$p_female[i]
  p_m = TableCountries$p_male[i]

  if (any(is.na(c(income, gpg, p_f, p_m))) || (p_m + p_f * (1 - gpg) == 0)) {
    male_income[i] = NA
    female_income[i] = NA
  } else {
    male_income[i] = income / (p_m + p_f * (1 - gpg))
    female_income[i] = male_income[i] * (1 - gpg)
  }
}
TableCountries$male_income = round(male_income, 2)
TableCountries$female_income = round(female_income, 2)
```

```
# Germany sectoral level
male_income = numeric(nrow(Germany_subgroup))
female_income = numeric(nrow(Germany_subgroup))
for (i in 1:nrow(Germany_subgroup)) {
  income = Germany_subgroup$income_eur[i]
  gpg = Germany_subgroup$gpg_2023[i] / 100
```

```
  p_f = Germany_subgroup$p_female[i]
  p_m = Germany_subgroup$p_male[i]
  male_income[i] = income / (p_m + p_f * (1 - gpg))
  female_income[i] = male_income[i] * (1 - gpg)
}
Germany_subgroup$male_income = round(male_income, 2)
Germany_subgroup$female_income = round(female_income, 2)
```

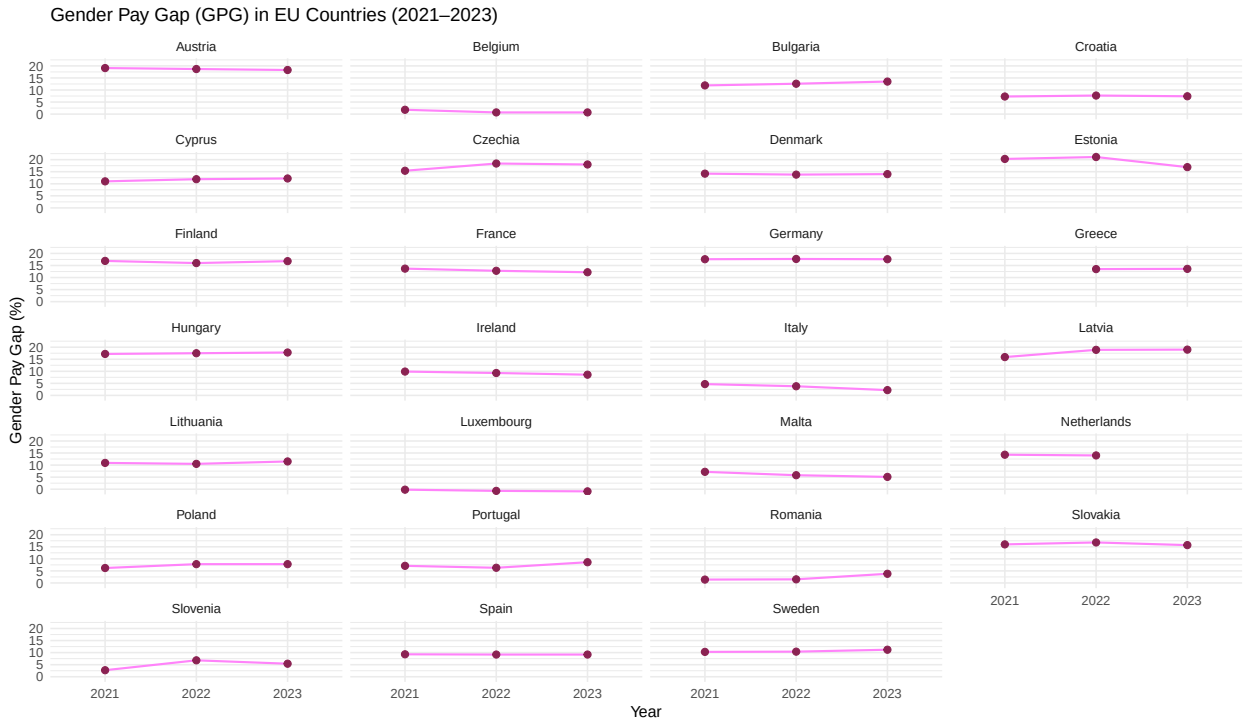
3.3 Visualize Temporal Variation

```
eu_countries = c("Austria", "Belgium", "Bulgaria", "Croatia", "Cyprus", "Czechia", "Denmark", "Estonia"

TableCountries_filtered = TableCountries %>%
  filter(country %in% eu_countries) %>%
  arrange(factor(country, levels = eu_countries))

ggplot(TableCountries_filtered, aes(x = year, y = gender_pay_gap, group = 1)) +
  geom_line(color = "orchid1", linewidth = 0.7) +
  geom_point(color = "violetred4", size = 2) +
  facet_wrap(~ country, ncol = 4, nrow = 7) +
  scale_y_continuous(limits = c(-1, 22)) +
  labs(
    title = "Gender Pay Gap (GPG) in EU Countries (2021-2023)",
    x = "Year", y = "Gender Pay Gap (%)"
  ) +
  theme_minimal(base_size = 11)
```

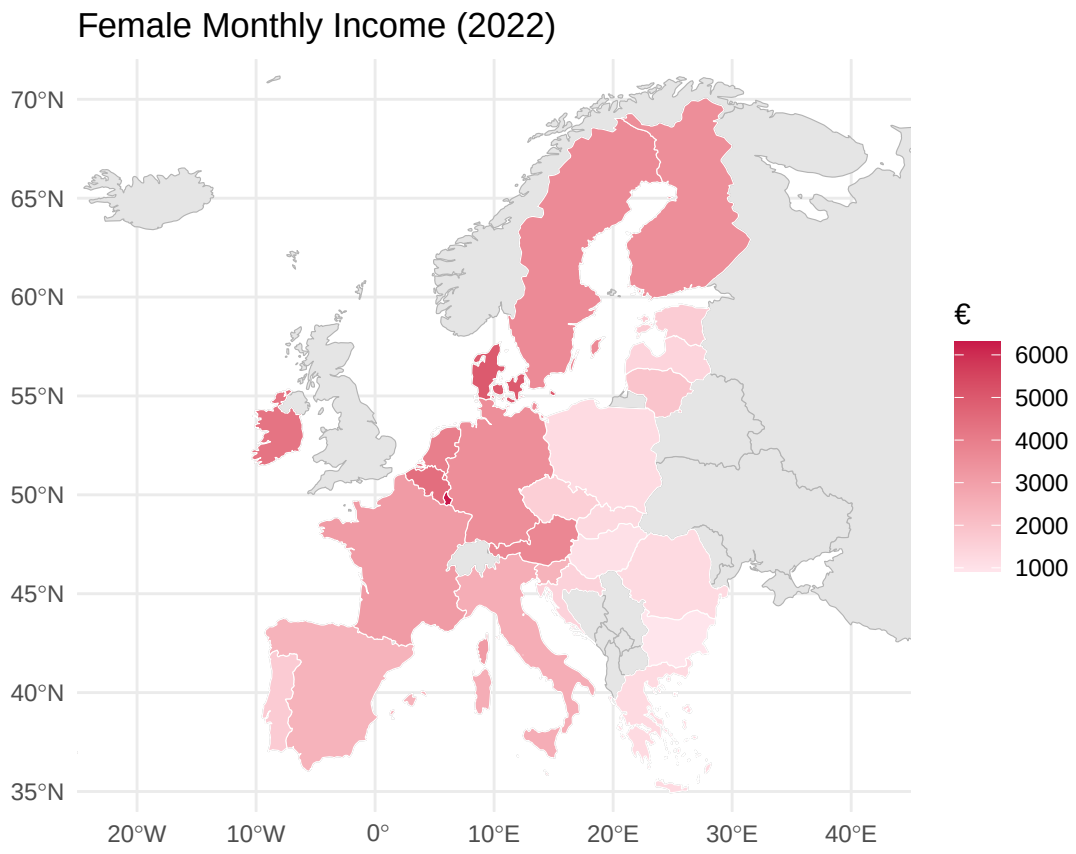
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



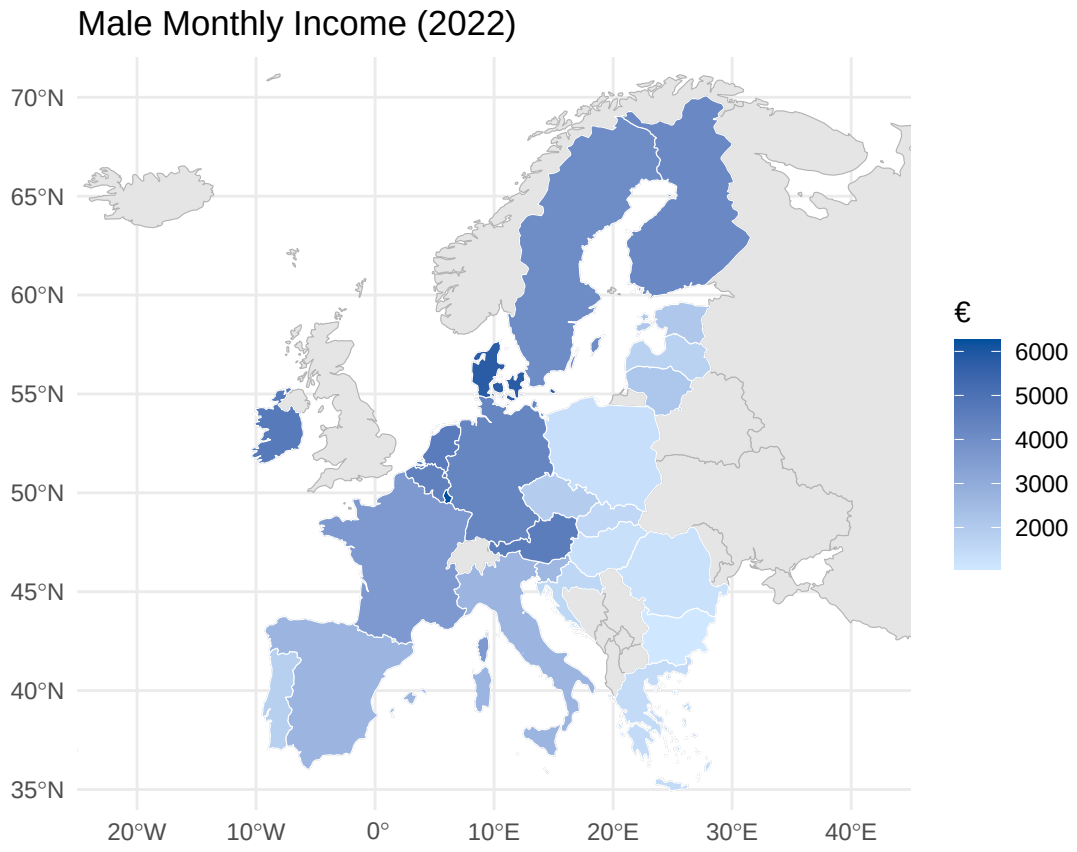
3.4 Visualize Spatial Variation

```
all_europe = rnaturalearth::ne_countries(scale = "medium", returnclass = "sf") %>%
  filter(region_un == "Europe")
eu_data_2022 = TableCountries %>% filter(year == 2022)
europe_map = all_europe %>%
  left_join(eu_data_2022, by = c("name" = "country")) %>%
  filter(!is.na(monthly_income))

ggplot() +
  geom_sf(data = all_europe, fill = "grey90", color = "grey70", size = 0.2) +
  geom_sf(data = europe_map, aes(fill = female_income), color = "white", size = 0.3) +
  scale_fill_gradient(low = "#FFE5EC", high = "#C9184A", name = "€") +
  labs(title = "Female Monthly Income (2022)") +
  coord_sf(xlim = c(-25, 45), ylim = c(34, 72), expand = FALSE) +
  theme_minimal()
```




```
ggplot() +
  geom_sf(data = all_europe, fill = "grey90", color = "grey70", size = 0.2) +
  geom_sf(data = europe_map, aes(fill = male_income), color = "white", size = 0.3) +
  scale_fill_gradient(low = "#D0E8FF", high = "#00509D", name = "€") +
  labs(title = "Male Monthly Income (2022)") +
  coord_sf(xlim = c(-25, 45), ylim = c(34, 72), expand = FALSE) +
  theme_minimal()
```



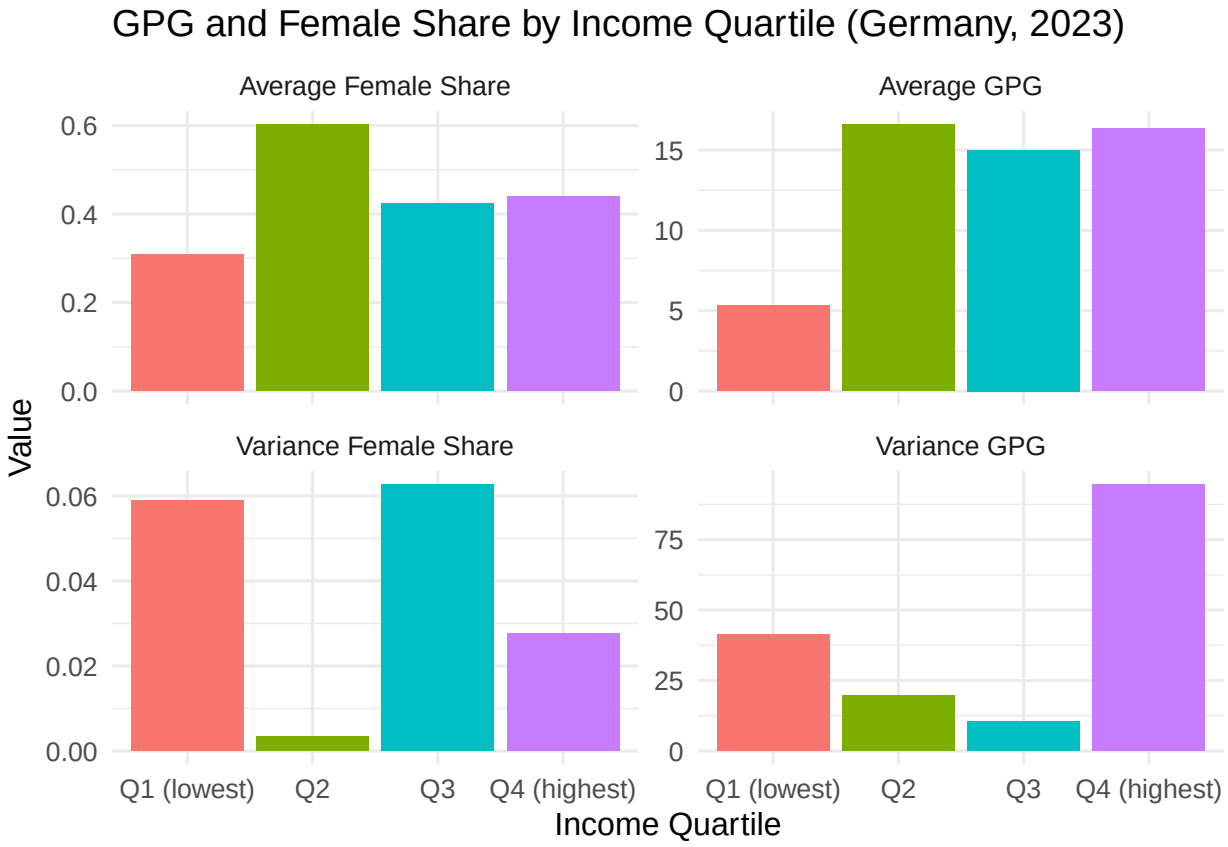
3.5 Visualize Sub-Population Variation

```
quartile_summary = Germany_subgroup %>%
  group_by(income_quartile) %>%
  summarise(
    avg_income = mean(income_eur),
    avg_gpg = mean(gpg_2023),
    var_gpg = var(gpg_2023),
    avg_p_female = mean(p_female),
    var_p_female = var(p_female),
    n_sectors = n()
  )

quartile_long = quartile_summary %>%
  select(income_quartile, avg_gpg, var_gpg, avg_p_female, var_p_female) %>%
  pivot_longer(cols = -income_quartile, names_to = "metric", values_to = "value") %>%
  mutate(metric = recode(metric,
    avg_gpg = "Average GPG",
    var_gpg = "Variance GPG",
    avg_p_female = "Average Female Share",
    var_p_female = "Variance Female Share"
  ))

ggplot(quartile_long, aes(x = income_quartile, y = value, fill = income_quartile)) +
```

```
  geom_col(show.legend = FALSE) +
  facet_wrap(~ metric, ncol = 2, scales = "free_y") +
  labs(
    title = "GPG and Female Share by Income Quartile (Germany, 2023)",
    x = "Income Quartile", y = "Value"
  ) +
  theme_minimal(base_size = 12)
```



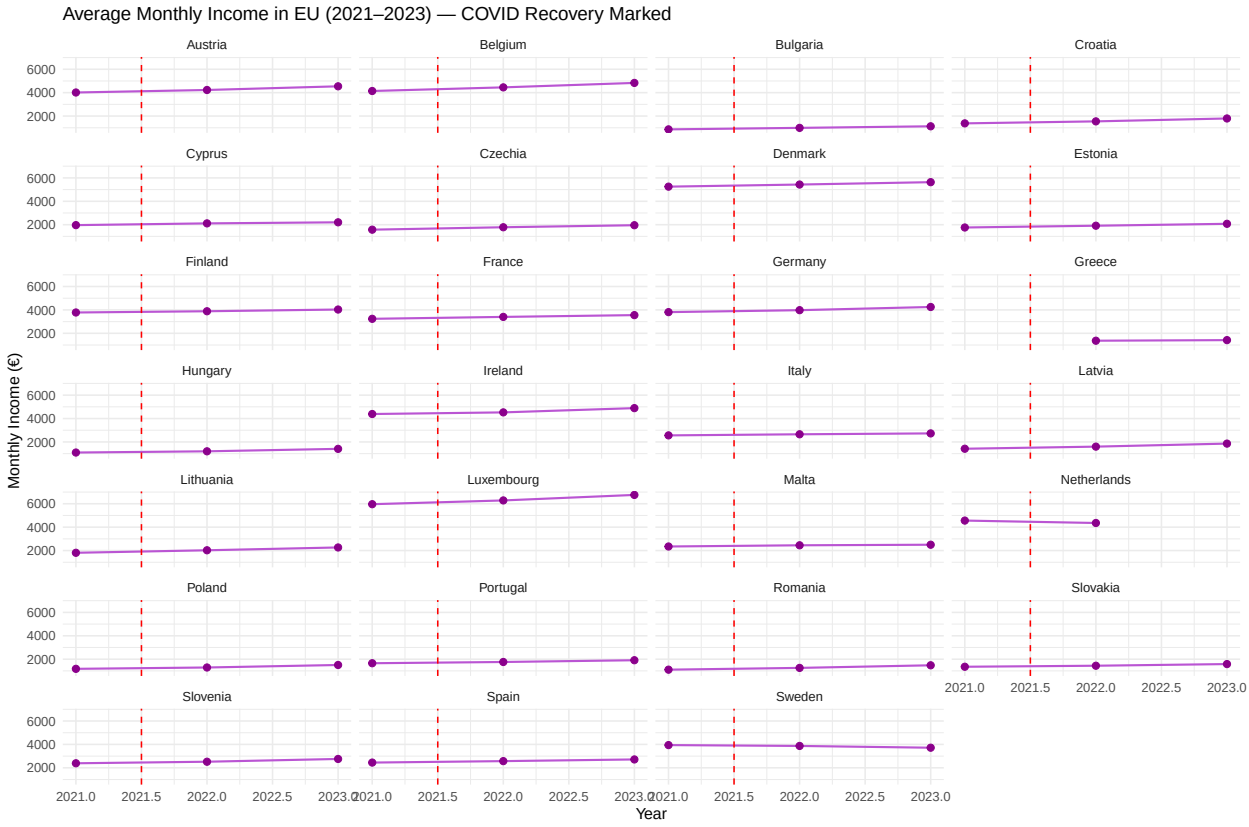
3.6 Event Analysis

```
filtered_data = TableCountries %>%
  filter(country %in% eu_countries) %>%
  mutate(year_numeric = as.numeric(as.character(year)))

ggplot(filtered_data, aes(x = year_numeric, y = monthly_income, group = 1)) +
  geom_line(color = "mediumorchid", linewidth = 0.7) +
  geom_point(color = "darkmagenta", size = 2) +
  geom_vline(xintercept = 2021.5, linetype = "dashed", color = "red", linewidth = 0.5) +
  facet_wrap(~ country, ncol = 4, nrow = 7) +
  labs(
    title = "Average Monthly Income in EU (2021-2023) - COVID Recovery Marked",
    x = "Year", y = "Monthly Income (€)"
```

```
  ) +
  theme_minimal(base_size = 11)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Part 4 - Discussion

4.1 Discuss Your Findings

WRITE THISSSSSSS

Part 5 - Reproducibility

5.1 Github Repository Link

THISSSSS

5.2 Reference List

THISSSS