

Data Exploration: Gender and World View

your name here

October 20, 2025

In this Data Exploration assignment, you will work with data from [Anduiza and Rico \(2024\)](#). It is panel data from Spain collected around the massive feminist protests of 2018 and 2019 and the subsequent rise of the far-right party Vox. The authors ask how levels of sexism changed following these events, and whether sexism increases far right voting. We will also use their data to examine whether different gender groups exhibit different political attitudes and reacted to these events differently.

If you have a question about any part of this assignment, please ask! Note that the actionable part of each question is **bolded**.

Data Details:

- File Name: `gender_panel.csv`
- Source: These data are from Anduiza and Rico (2022).

The table below documents the variables available in the data. See the file `data_codebook.pdf` for more detailed descriptions of variables.

Variable Name	Variable Description
<code>codpanelista2</code>	Respondent unique identifier
<code>wave</code>	Panel survey wave (9=2017, 10=2018, 11=2019, 12=2020)
<code>sex</code>	1 = Male, 2 = Female
<code>age</code>	Age
<code>education</code>	Highest level of education completed (1-11)
<code>hhincome</code>	Income
<code>livingpartner</code>	Live with my partner (0=No, 1=Yes)
<code>lrself</code>	Political ideology (0=Extremely left, 10=Extremely right)
<code>voteintentionspain</code>	The party the respondent would vote for if the election were held tomorrow (1-55, 23=Vox)
<code>polintr</code>	How interested are you in politics (1=Very much, 2=Somewhat, 3=A little, 4=Not at all)
<code>femindexa-i</code>	9 questions measuring modern sexism. See Question 1.
<code>authoa-d</code>	4 questions measuring authoritarianism. See Question 1.
<code>populisma-f</code>	6 questions measuring populism.
<code>immigeco</code>	The effect of immigrants on the economy (0=Very negative, 10=Very positive)
<code>immicult</code>	How important it is for coexistence that immigrants have the same culture and customs as people here (0=Not at all, 10=Very important)
<code>femstrike</code>	Went on strike on Women's Day (1=Yes, 2=No)
<code>femdemonstrate</code>	Participated in a Women's Day demonstration or rally (1=Yes, 2=No)
<code>feminfo</code>	Sent information or made comments inviting others to participate in Women's Day rallies (1=Yes, 2=No)

Variable Name	Variable Description
femtalk	Talked about these activities on Women's Day with people around me (1=Yes, 2=No)

Question 1

Part a

The survey measures respondents' modern sexism (attitudes denying gender discrimination and rejecting efforts to rectify gender inequality) using 9 survey questions (`femindexa` to `femindexi`). **Create one new variable that aggregates each person's responses to these 9 questions. You will use this new variable in your analyses of sexism below.** Hint: The authors take the following steps: (1) reverse-code some variables so that a higher value always denotes higher sexism, (2) average responses in 9 columns, and (3) rescale the average so that it ranges between zero and one. See `data_codebook.pdf` to determine which variables need to be reverse-coded. Below we provide an example using authoritarianism.

The survey also measures authoritarianism that we covered last week, but using 4 questions based on childrearing values, questions different from last week. The code below creates one new variable that aggregates each person's responses to these 4 questions.

```
df <- df |>
  mutate(across(starts_with("autho"), ~ . == 2)) |>
  # reverse-code
  mutate(across(c(authoc), ~ 1 - .)) |>
  # average
  rowwise() |>
  mutate(authoritarianism = mean(c_across(starts_with("autho")), na.rm = TRUE)) |>
  ungroup() |>
  # rescale
  mutate(authoritarianism = rescale(authoritarianism, to = c(0, 1)))
```

Part b

What are the distributions of modern sexism and authoritarianism among survey respondents? Do men or women have higher levels of modern sexism and/or authoritarianism? Make density plots per gender. Conduct statistical tests on the gender differences.

Part c

Let's take a step back. Look back at the survey questions in `data_codebook.pdf`. **Do you think these 9 questions measure politically important dimensions of sexism? How about the 4 questions on authoritarianism? Do you think these childrearing questions are better at measuring authoritarian attitudes than last week's Aggression-Submission-Conventionalism scale? If the paper's goal was to study gender differences in sexism or authoritarianism (note that this was NOT the goal of Anduiza and Rico (2024)), do you think these measures were appropriate for such purpose? Do you think other measures might have produced different conclusions?** There are no definite answers. Feel free to think about it and share your opinions.

Question 2

Part a

The dataset for this week is unique in that this is *panel* data. This means that the same set of respondents were observed repeatedly over time, in our case, 4 times across 4 years. **Make a time-series plot showing the changing levels of modern sexism across years. Plot the trend per gender. In the plot, add**

a vertical line in between wave 2017 and 2018 to indicate when Women's Day rallies happened. You can use `geom_vline()` and `annotate()` to make this clear. Are there any changes across time? The authors argue that the salient feminist mobilization initially suppressed sexism but provoked a backlash afterwards. They also claim that the subsequent rise of far-right party Vox normalized sexist attitudes. Does the time trend seem to be consistent with this claim? Is there any gender differences in temporal trends?

Part b

We can look at aggregate time trends using panel data, but the utmost benefit of it is that we can look at the time trend per individual. Pick several respondents that you find interesting based on their demographic or political profiles. Plot how their political ideology `lrsel` changed over time. Do the trends differ across people depending on their characteristics? If you find any interesting patterns, check whether the pattern is generalizable, i.e. similar patterns are found for many respondents with the same profile. For this purpose, you can plot the group-level average trend. Feel free to pick other variables that you find interesting and explore further.

Question 3

Let's take an initial look at how modern sexism relates to far right voting. Create a variable indicating whether a respondent said they would vote for Vox if an election happened tomorrow (`voteintentionspain`). Conduct a bivariate regression using the variable that you created for modern sexism. Interpret your results. If you'd like, extend the problem by reporting the 95% confidence interval for the `modern_sexism` coefficient. Is the relationship statistically significant?

Hint: Once again, the `lm()` function is used to calculate linear regressions in R. [Here](#) is a guide to linear regression in R that may be helpful.

Question 4

The model in the previous question was very simple; we modeled the vote choice only as a function of modern sexism. In reality, the relationship is probably more complicated - other demographic and political characteristics of respondents can affect both their vote choices and their levels of modern sexism. Thus, we would like to control for these confounding influences by including those variables as control variables. Look at the other variables available in the data and find one or more that you think may also be related to both far right voting and modern sexism. Explain why you think so, then add the variable(s) to the right side of the regression (as explanatory variables). You can use `stargazer()` function to present the results in a cleaner form. Interpret what you find.

Question 5: Data Science Question

Is modern sexism more likely to increase far right voting among men than women? Estimate a regression of the following form: $\text{voteVox} \sim \text{modern_sexism} + \text{sex} + \text{modern_sexism}:\text{sex}$, where `modern_sexism:sex` is the interaction between modern sexism and respondents' sex. Compare this to the same model but without the interaction term. Interpret your results.

In the social sciences, we use interaction terms in regressions to capture heterogeneous effects. As an example of how to implement and interpret this type of model, suppose we wanted to understand the relationship between education on the one hand (as the outcome variable), and age and gender on the other hand (as explanatory variables). We might think that the effect of age on education depends on whether you're looking at men or women. Maybe for men, age has no effect on education, but for women, there is a negative effect, as older women were discouraged or barred from seeking higher education. To assess whether this is true, we can use an interaction between gender and age. You can model this in R using this formula in the `lm()` function: `education ~ age + female + age:female` (supposing gender is coded into a binary variable `female`). Here, `age:female` is what creates the interaction.

Let's say that we ran this regression in R and found that the model looks like this: $education = 1.5 + 0.005 * age + 0.01 * female + -0.4 * age * female$. Here, the coefficient on **age** is 0.005, 0.01 on **female**, and -0.4 on the interaction between the two. Without an interaction, to interpret the coefficient on **age**, we would say the effect of **age** on **education** is 0.005. However, the interaction term modifies that relationship, i.e. the effect of **age** on **education** now depends on gender.

To see this, we must plug in values for **female** and **age**. When **female** = 0, then the interaction term vanishes, and then the effect of **age** on **education** is 0.005. In other words, for non-women, there is a very small relationship between age and education. Now plugging in **female** = 1, the effect of **age** on **education** becomes 0.005 (the coefficient on age) + -0.4 (the coefficient on the interaction) = -0.395. In other words, the effect of age on education among women is negative.

Interpreting the effect of gender is a bit more complicated, and in this case is nonsensical. To do so, we set **age** = 0 (which doesn't make a ton of sense) to find that the effect of **female** on **education** is 0.01 when **age** = 0. Always pay attention to whether the coefficients you're focusing on are even substantively meaningful.

Question 6

The data contain many interesting variables. **Think about hypotheses related to gender that you want to test using this data. Use plotting functions and multivariate regressions to test your hypotheses. Interpret your findings. Does the data support your hypotheses?** Looking forward to seeing your creative responses.