Final Project example

Ramses Llobet

February 25, 2024

Data cleaning and preliminaries (not be included in your submissions)

First, ensure to clean and prepare your data for analysis. Remember **not to include this section** in your final project submission. Treat your final project as if it were a report for a consulting firm or a submission to a scientific journal. Therefore, avoid displaying code or unnecessary computations. Instead, focus on presenting figures, tables, and statistics that are essential for addressing your research question, conducting analysis, and drawing conclusions. Additionally, **refrain from displaying any code**. Only present well-formatted figures, tables, and results. Ensure that your general chunk options include the argument `echo = FALSE`.

I will grade your assignments based on:

- **Well-written content** (proofread), including research questions, hypotheses, analysis, interpretation, discussions, and conclusions, with accurate interpretations. Misinterpreting results or statistical tests will affect your final grade.
- Adherence to the structure outlined in this template, following the delineated sections:
  - Introduction
  - Theory and hypotheses
  - Data and visual analysis
  - Statistical analysis
  - Conclusions
- Presentation of **formatted figures and tables** with informative labels and rounded numbers. I understand the challenges of positioning tables or figures and will be forgiving in this regard, as fixing their placement in your PDF output can be challenging for RMarkdown beginners. I will provide code and tips in this template.

Below, I've addressed the cleaning of this project. The code chunk has the option `include=FALSE` to **prevent** the output from this data cleaning part being displayed in the final printed PDF version. **You must not show how you clean your data.** However, in the *data and methods* section, you must report and explain any significant modifications or transformations of the raw data, such as the percentage of missing data and the extent to which your sample size decreased after cleaning.

Please note that we will be using the `transphobia_all.csv` dataset, which can be found in the application of *Brookman and Kalla's (2016)* paper.

A note on formatting tables and using stargazer()

In your final project, ensure you present formatted tables. I recommend using the `stargazer()` function, but you can also explore other options like `KableExtra` or `pander`.

When using `stargazer()`, consider these arguments and chunk options to enhance presentation:

- Set `header=FALSE` to eliminate a redundant message from the function author.
- Adjust `column.sep.width = "1pt"` to reduce column separation width.
- Utilize the `title` argument to name the table.
- Specify `digits = 2` or your preferred number to limit displayed decimals.
- Choose `type = text` for console printing and `type = latex` for PDF output in LaTeX. Remember to set the code chunk option `results='asis'` for proper rendering.

If a `stargazer` table in the PDF **floats** away from its code **position**, it is acceptable. Don't worry excessively about table placement. If a table becomes overly wide, consider splitting it into two. Focus on reporting important variables and results, and avoid tables with no significant information.

You can see how I have programmed the below tables in the original `project_example.Rmd`.

Introduction

This is the first section, it introduces your project topic and its significance. It outlines **why the topic is important** and what insights it can offer. Ideally, you should present it as a puzzle, and your analysis is going to contribute to one piece to "solve" this puzzle. Additionally, you should briefly outline your research question, units of analysis, and other contextual factors such as the year of the data and the number of countries, regions, or respondents involved. This section should be concise, spanning no more than 2 or 3 paragraphs.

Theory and hypotheses

This section begins by revisiting the **research question**: *"Why do people hold prejudices towards other groups or communities?"* and *"What factors can contribute to the reduction of outgroup prejudice?"* Research questions often start with **"Why"** or **"What"** and may be follwed by narrower question that explore the same relationship.

After introducing the research question, , a brief literature review should be included, summarizing previous research and theories related to the project topic. This review helps in narrowing the analysis, selecting statistical controls, and possibly adding additional hypotheses. No particular citation format is prefered.

Towards the end of this section, narrow hypotheses should be provided, addressing a causal relationship of interest ($X \rightarrow Y$) based on the research question. It's crucial to explain the underlying causal mechanism ($\rightarrow$) between the predictor X and the outcome Y.

In this project, the causal assumption/mechanism ($\rightarrow$) is that *communication* and *information* (X) can influence beliefs and prejudices towards an outgroup (Y). This psychological mechanism suggests that outgroup prejudice stems from limited empathy and interaction with specific minorities and their struggles. Thus, providing accurate and credible information about these minorities should bridge the gap between subjective and objective knowledge, resulting in reduced outgroup prejudice. In other words, we expect a **negative relationship** between providing more sensible information about outgroups and outrgroup prejudice.

Based on this causal assumption, two hypotheses are considered:

- H_0: **Conversations** promoting the active adoption of others' perspectives do not reduce prejudice. ($\beta_{treatment} = 0$)
- H_1: **Conversations** promoting the active adoption of others' perspectives reduce prejudice. ($\beta_{treatment} < 0$)

Typically, the null hypothesis H_0 is implicit in the alternative H_1. However, for clarity, both are stated here.

Data and visual analysis

Data

In this section, report your data collection and research design.

Begin reporting your sample size and rellevant data cleaning procedures. Discuss whether the design is **experimental** or **observational**, detailing treatment assignment and potential confounding. Report any variable transformations made. For instance, I inverted the

scale of `therm_trans_t1` to create `prejudice_trans`, to reflect a negative association with treatment, as hypothesized in H_1.

After detailing each variable and data, provide well-formatted tables with descriptive statistics, such as a correlation matrix and measures of centrality and dispersion. I will use the `stargazer()` function to present formatted tables, but you can use any other functions.

Table 1: Descriptive statistics

Statistic	Mean	St. Dev.	Min	Median	Max
prejudice_trans	42.71	28.51	0	50	100
treat_ind	0.48	0.50	0	0	1
age	48.28	17.63	17	50	90
female	0.59	0.49	0	1	1
democrat	0.47	0.50	0	0	1
voted_gen_12	0.78	0.42	0	1	1

Table 2: Correlation matrix

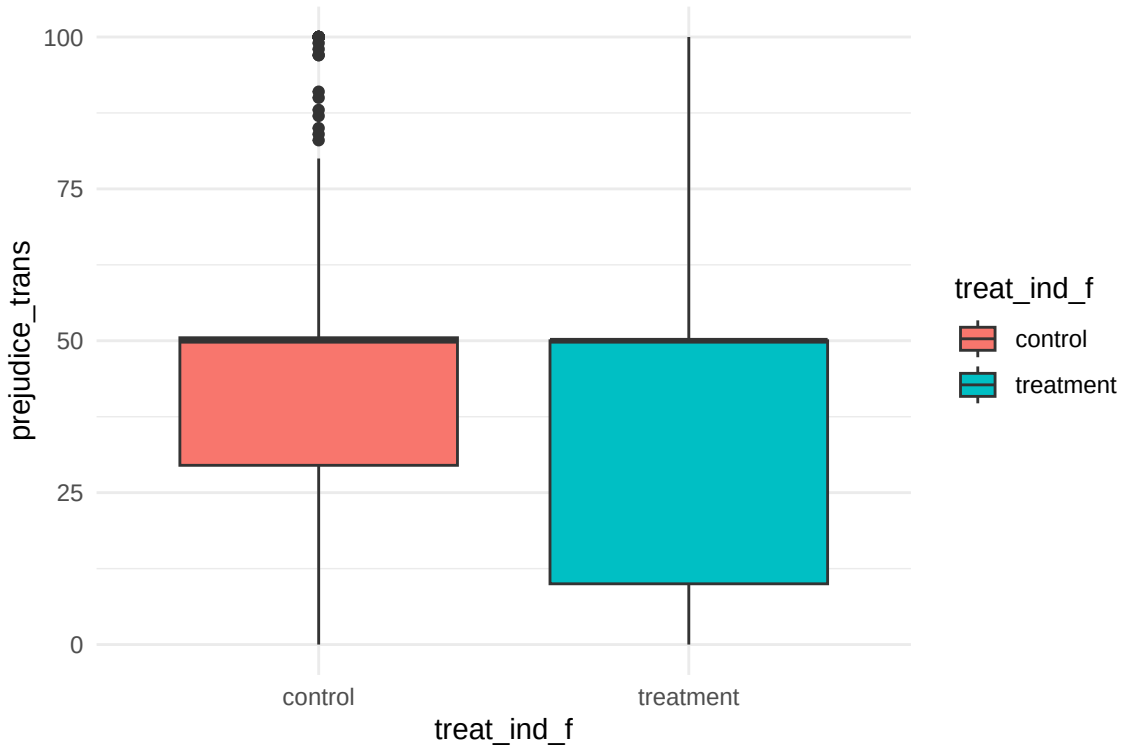
	prejudice_trans	treat_ind	age	female	democrat	voted_gen_12
prejudice_trans	1	-0.12	0.17	-0.13	-0.08	0.06
treat_ind	-0.12	1	0.02	-0.02	0.03	-0.11
age	0.17	0.02	1	-0.06	0.02	0.34
female	-0.13	-0.02	-0.06	1	0.12	0.01
democrat	-0.08	0.03	0.02	0.12	1	0.12
voted_gen_12	0.06	-0.11	0.34	0.01	0.12	1

Present the correlation coefficients and analyze them. In this instance, the treatment variable `treat_ind` exhibits a negative correlation of -0.12 with the outcome `prejudice_trans`. Additionally, `age` demonstrates a positive correlation of 0.17 with prejudice, while `female` displays a negative correlation of -0.13. The positive correlation between `age` and `prejudice_trans` might indicate cohort effects, suggesting that older respondents are more likely to harbor prejudices toward the trans community.

Exploratory data analysis

In this section, conduct a **visual analysis** of the variables of interest. Begin by ploting correlations among the main variables of interest, followed by exploring the data structure and relationships using visuals.

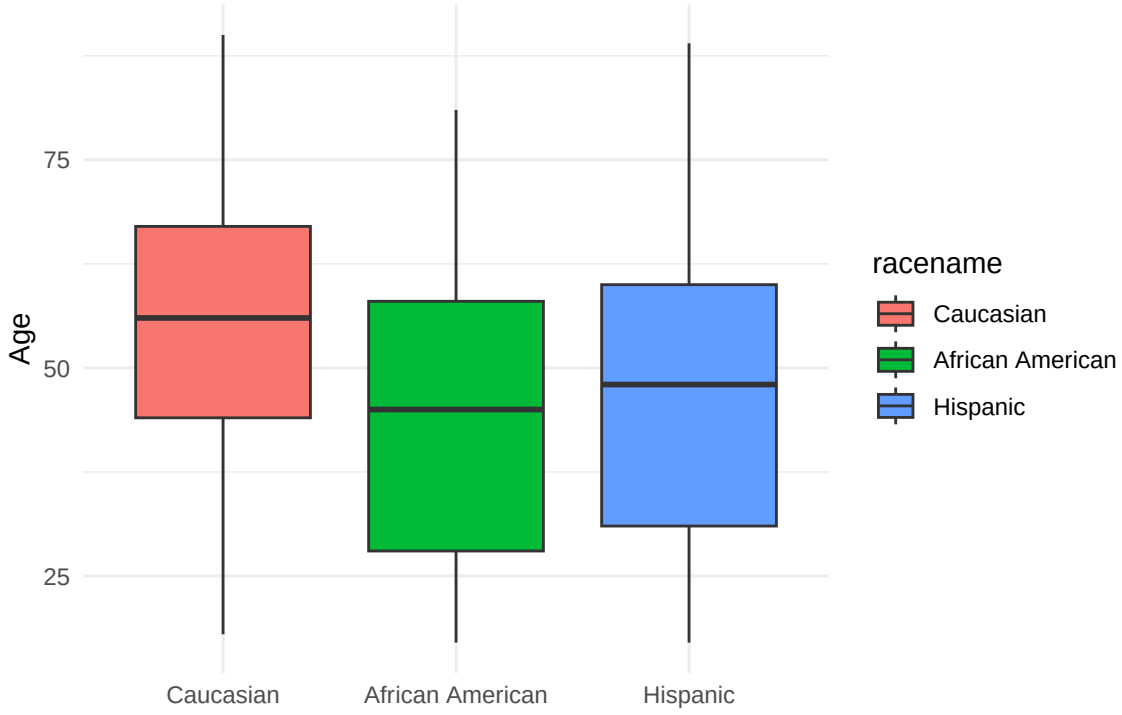
Choose the most **meaningful visuals** to include in your final project and refine them accordingly. Additionally, in your final projects, **provide a narrative** explaining your selection of specific visuals and how they inform your model selection to test your hypotheses.



The above is an example of a **bad visualizaiton**, which you **should not report**!

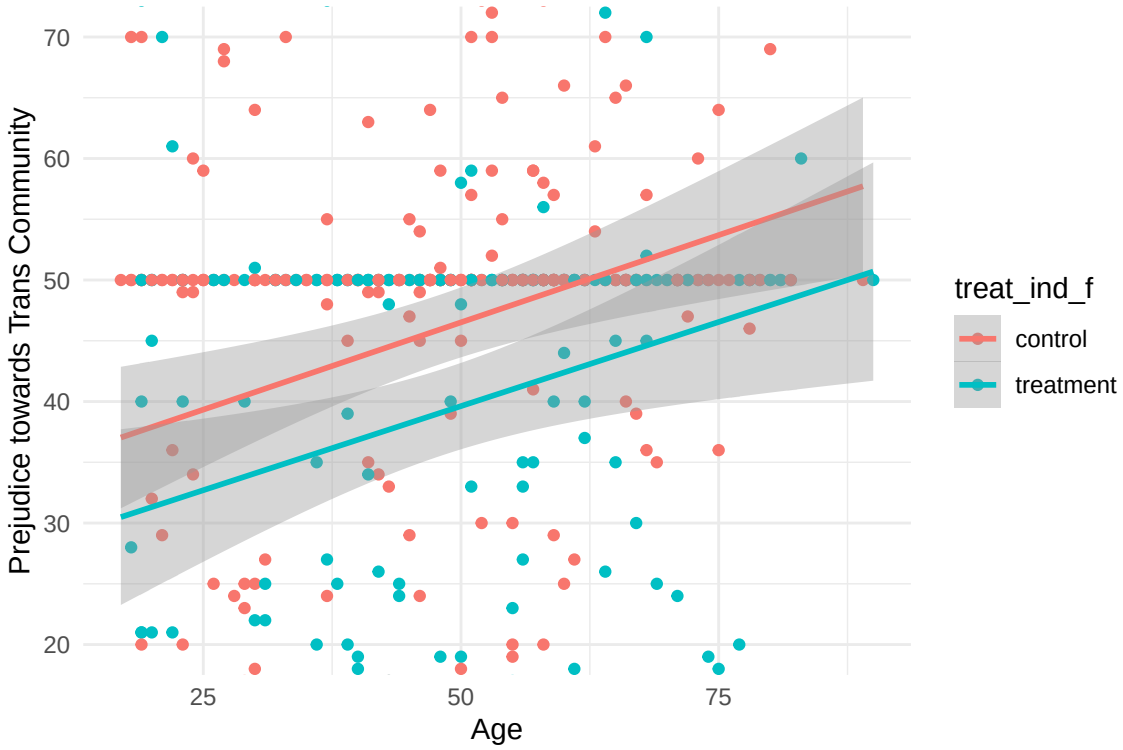
The provided boxplot illustrates a **skewed distribution** of prejudice, which compromises its visual effectiveness. The skewed distribution results in the **median** and the **third** quartile aligning, obscuring the box's appearance. However, the plot offers valuable insight: the distribution between the median and the first quartile is wider in the treatment group compared to the control group. This suggests that more respondents from the treatment group are dispersed across the lower half of the distribution, indicating lower levels of prejudice. You can transmit the same insight by simply reporting a **differences in means** of prejudice between the treatment and control group.

We saw that **age** is positively correlated with prejudice. We can look at the distribution of age as a function of the respondent's race.



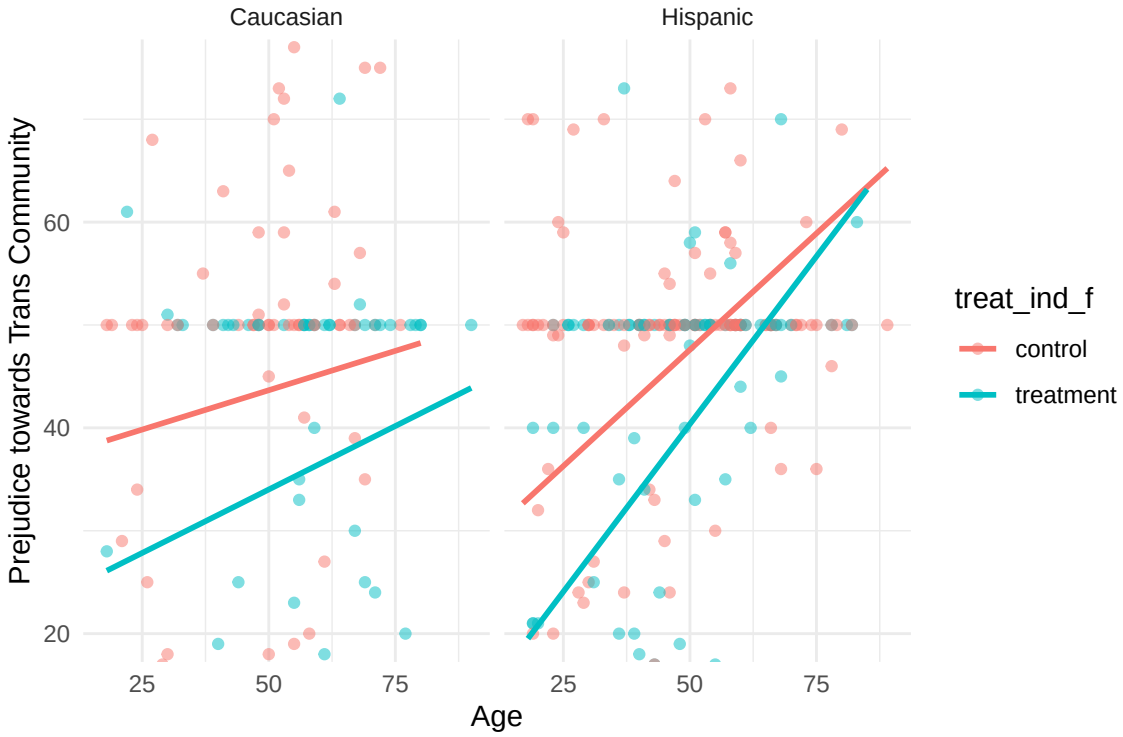
The boxplot indicates that the median age of Caucasian respondents is higher compared to African American and Hispanic individuals. Given the importance of age as a control variable and the likely correlation between race and outgroup prejudice, we should include the "race" factor in the statistical analysis to account for its influence on the treatment's marginal effect.

We can now explore linear relationships between age and prejudice, conditioning on the treatment group to see whether treatment effects hold controlling for age.



Age correlates positively with prejudice. However, upon considering treatment status, we observe a lower initial level of prejudice in the treatment group compared to the control. This difference implies that, on average, respondents receiving the treatment exhibit lower prejudice levels, even after accounting for age.

To delve deeper, we'll introduce another factor: race. Specifically, we'll focus on differences between Caucasian and Hispanic respondents in adjusted prejudice levels by age.



For simplicity, I removed the confidence intervals by setting the argument `se=FALSE` in `geom_smoth()`. Moreover, I used `facet_wrap()` as a function of `racename` to split the visual in two subplots, one doing the regression for the Caucasian sample and the other for Hispanics.

Statistical analysis

After completing the exploratory data analysis and conducting preliminary regressions, we will select the **model specification** with the relevant predictors, including necessary controls. In your liner models, you shoudl only inlude interactions of interest, the main predictor/treatment, and only controlling for confounders Z. That is, variables that follow this relationship $X \leftarrow Z \rightarrow Y$.

Recall:

- H_0: Conversation encouraging the active adoption of others' perspectives **does not** reduce prejudice. ($\beta_{treatment} = 0$)
- H_1: Conversation encouraging the active adoption of others' perspectives reduces prejudice. ($\beta_{treatment} < 0$)

Linear regression analysis

Our strategy will involve presenting the model **progressively**, beginning with a bivariate regression between the independent and dependent variables. We will then gradually incorporate relevant controls until we present the full specification. We aim to limit the number of models to a maximum of four.

The results presentation should consistently provide evidence directly related to our **hypothesis** and assess whether this evidence **substantially changes** with the inclusion of controls.

Again, I will be using `stargazer()` function to report the estimated models in table.

Table 3: Regression analysis of treatment on prejudice

	prejudice_trans			
	Model 1	**Model 2**	**Model 3**	**Model 4**
treat_ind	−6.700***	−6.874***	−7.131***	−7.204***
	(2.331)	(2.297)	(2.269)	(2.249)
age		0.282***	0.311***	0.379***
		(0.065)	(0.066)	(0.070)
female			−8.035***	−7.389***
			(2.333)	(2.325)
racenameAfrican American			10.608***	14.296***
			(3.278)	(3.463)
racenameHispanic			6.137**	4.982*
			(2.807)	(2.802)
democrat				−7.671***
				(2.594)
voted_gen_14				−5.142**
				(2.515)
Constant	45.935***	32.397***	30.220***	33.074***
	(1.617)	(3.509)	(4.474)	(4.501)
N	592	592	592	592
R-squared	0.014	0.044	0.075	0.096
Adj. R-squared	0.012	0.041	0.067	0.085

***p < .01; **p < .05; *p < .1

Results discussion

After the analysis, you should discuss the following:

- Does the evidence support rejecting the **null hypothesis** (H_0) in favor of the **alternative hypothesis** (H_1)?

- Is this evidence **consistent** across different model specifications and **significance levels**?
- What **future research** suggestions can enhance our understanding of these findings and further investigate the research question in your project?

Conclusion

In the conclusion, succinctly summarize the analysis conducted in 2 or 3 paragraphs. Reflect on your original research question and hypotheses, detailing the data and research design employed. Evaluate the results obtained from your analysis and assess whether they align with your initial hypotheses. Importantly, **refrain from introducing new information**; instead, offer a concise recapitulation of the preceding sections and reported results.