US General Society Survey Analysis[*]

Cristina Burca, Yan Mezhiborsky, Sakura Noskor

March 16, 2024

Abstract

Some abstact

Table of contents

[*] Code and data are available at: repository

Introduction

This research examines the voting patterns in the 2016 and 2020 US Presidential elections. We will be looking at data collected by the Cooperative Election Study and accessed through the Harvard University Database **cite**. The anaylsis is based on a representative sample of 61,000 American adults, which provides detailed information about each individuals gender, birth year, race, registered state, employment, education loans, immigration status, dual-citizenship, religion, and 2016 and 2020 Presidential vote. The goal of this study is to use relevant variables from the electoral data to investigate patterns, trend and predictions regrading American electoral preferences from 2016 and 2020.

Data

Data Used

This paper was modeled with the help of R (R Core Team 2023) along with other useful packages like tidyverse (Wickham et al. 2019a) (which includes graphing functions like ggplot2), patchwork (Pedersen 2024). There are parts of the code which were guided by Rohan Alexander's Telling Stories with fire (Wickham et al. 2019b) chapter 13 section 13.2.2.

Variables inspected

Starting off, we examine the columns 'votereg' and 'voted_for'. They represent the number of persons that registered to vote and which candidate they voted for in 2020, respectively. We filtered out the rows with a 'votereg' value of 2, which indicated unregistered voters, to focus exclusively on individuals who were registered to vote. We then focused on the 'presvote16post' variable, which reveals the candidates Americans voted for in the 2016 United States Presidential Election. This is an important variable as it enables us to assess whether American citizens were satisfied with the service that the previous government provided. Next we look at 'gender' as well as 'employment'. Both 'gender' and 'employment' shows us if there is a correlation between certain parties views versus the demographic they represent. 'Gender' contains 2 values (male and female) while, employment has 9 values; full time worker, part time worker, laid off, unemployed, retired, permanently disabled individual, Homemaker, Student or Other. We also explore the variable 'immstat' which represents the immigration status of the of individual represented by one of the following: immigrant and citizen, immigrant not citizen, born in US, but parent(s) immigrant, parent and I born in US but grandparent(s) immigrant, or all born in US.

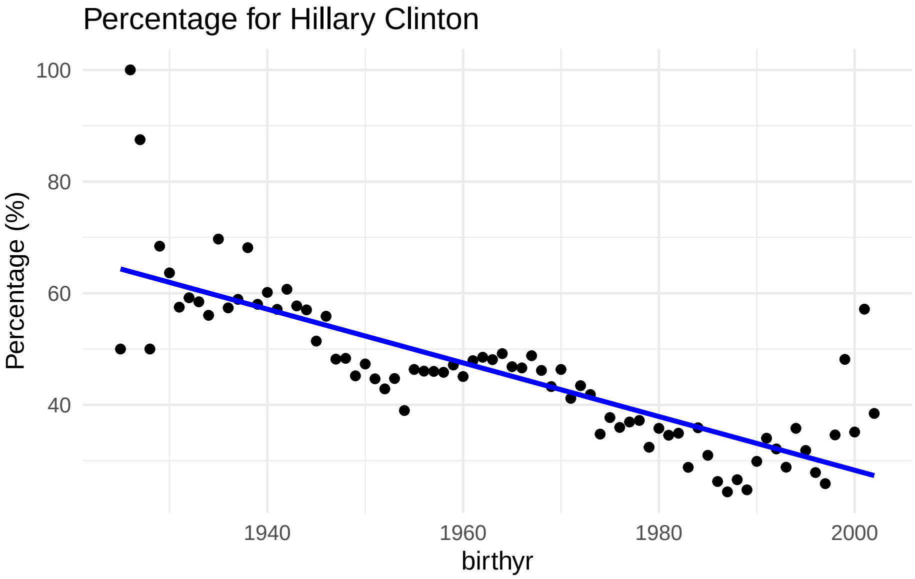
The Destination to Reach with the Data

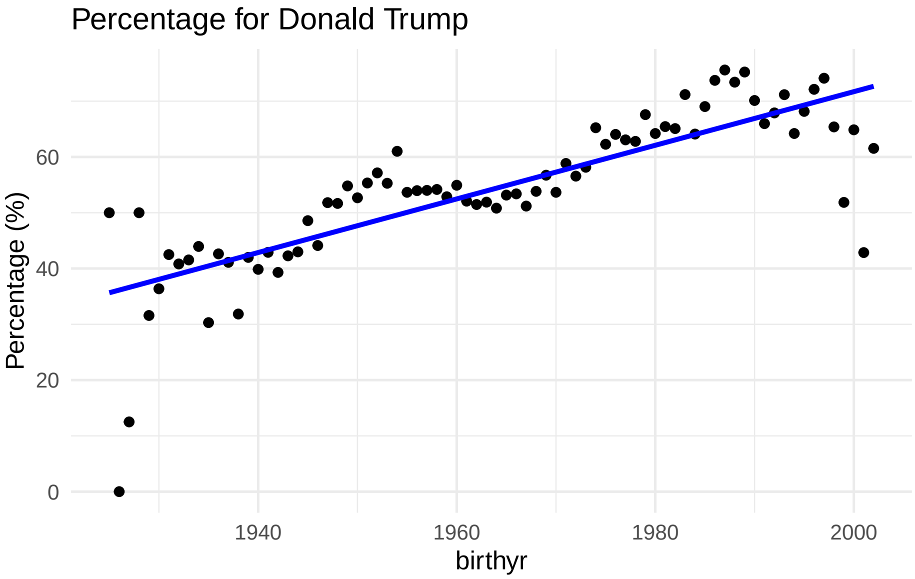
There could have been many other similar data sets that could have been used for this project for example we could have chosen to look at the census and election data for Canada. However, our group decided that because part of the analysis was done in Wickham et al. (2019b), there were still many other variables that we could explore as we dive further into the 2020 presidential election and try to interpret if there are any correlations between the variables and the result. Our team found it interesting to see all the variables that were collected by the US government and the correlations we saw during the analysis process; where there most definitely was a positive correlation between each variable and the outcome of the votes. Although we are analyzing the 2020 election that has already taken place, the analysis we do in the later sections are believed to apply to the 2024 elections happening this year. This is enough reason for us and the reader to dive into the patterns that exist with this large data set.

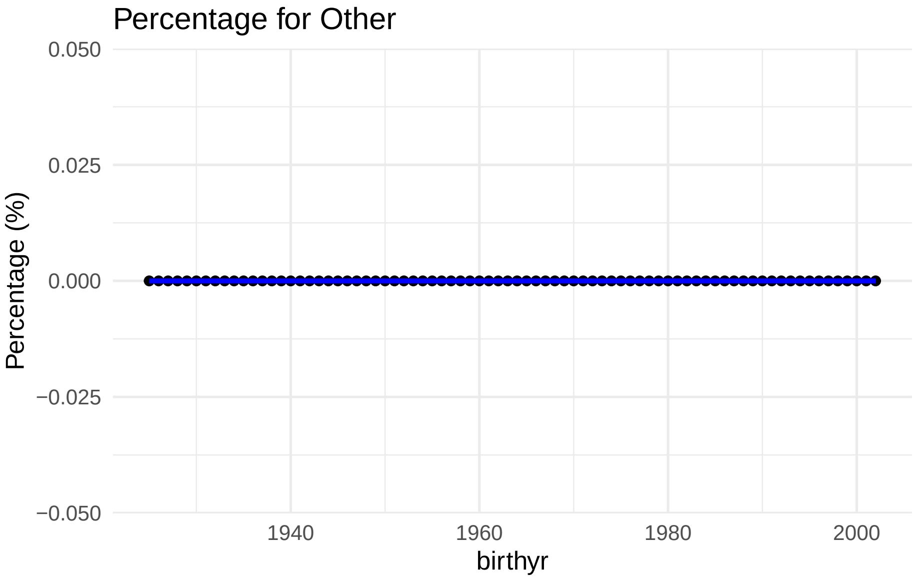
Models

Model 1

Initially, we filter the data to remove to exclude individuals who did not vote for Hillary Clinton or Donald Trump, due to the minimal votes for Gary Johnson, Jill Stein, Evan McMullin, and all other candidates which are insignificant to this paper. We first analyze the birth year of voters.


Percentage for Hillary Clinton
Percentage (%)
100
80
60
40
1940
1960
1980
2000
birthyr


Percentage for Donald Trump
Percentage (%)
60
40
20
0
1940
1960
1980
2000
birthyr



Model 2

In this model, we conduct an examination of the relationship between voters' birth year and their gender for the 2016 and 2020 Presidential elections. This analysis is visualized by plotting a histogram that separate female voters on the left and male voters on the right, with voters' birth years measured along the x-axis, which ranges from 1925 to 2002. The y-axis quantifies the voter turnout for the year. For clarity ad symbolic representation, the colour blue was chosen to represent the Democratic candidates –Hillary Clinton for the 2016 election, and Joe Biden for the 2020 election, while red was chosen to the Represent the Republican candidate, Donald Trump, who sought the presidency in both terms. Figure 1 present the distribution of votes in 2016, and Figure 2 presents the data from the 2020 election.

We notice in Figure 1 the graph displays two distinct high points for the

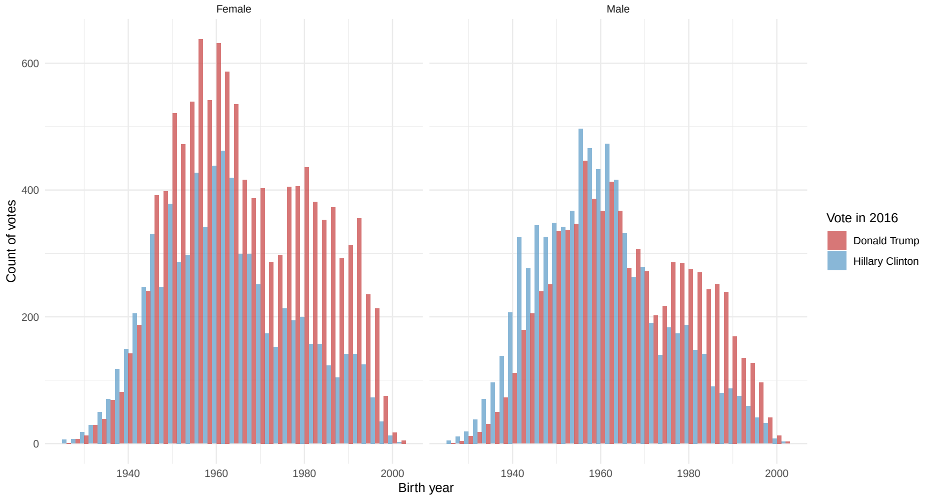


Figure 1: Logistic regression of 2016 US Presidential votes comparing parameters of gender and immigration status

number of female Republican voters, with the peak around 1960 being the most pronounced, followed by another around 1980. There are similar peaks in the graph for female Democratic voters, but with the overall count being considerably lower. We also notice that the Democratic party received slightly more votes from the older demographic, where as the younger demographic greatly preferred the Republican party. In comparison, the graph displaying the male votes has a more balanced distribution, and similar to the female voters, the older and younger demographic preferred the Democratic Party and the Republican Party, respectively. It is worthy to mention that there is a higher count of women than men.

We notice in Figure 2 that the graphs bear a strong resemblance to the distributions of the 2016 votes depicted in Figure 1, but with the parties reversed on the graph. We are interested in the ratios of these graphs.

We analyze the ratio of Democratic to Republic votes in 2016 in Table 1

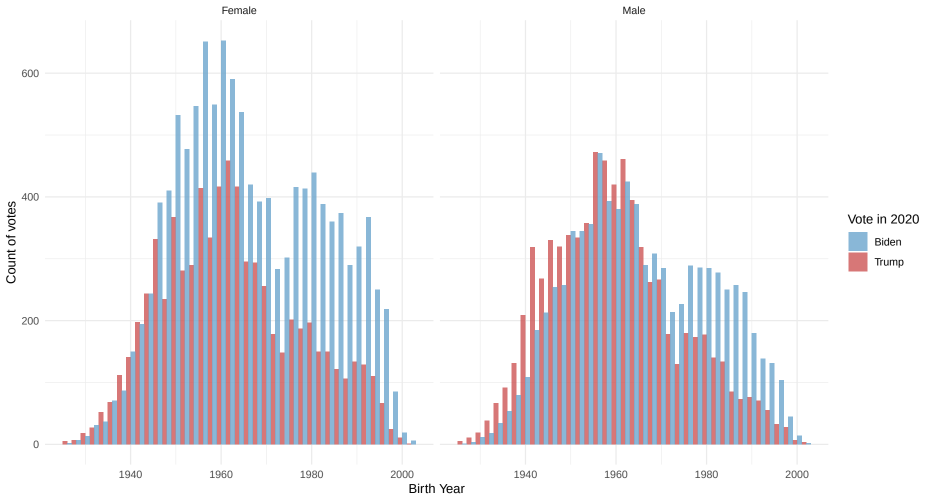


Figure 2: Logistic regression of 2020 US Presidential votes comparing parameters of gender and immigration status

Table 1

Table: Ratio of 2016 Votes for Clinton to Trump by Birth Decade (Females)

Birth Decade	Total Clinton Votes	Total Trump Votes	Ratio
1930	285	231	1.23
1940	1179	1360	0.87
1950	1730	2712	0.64
1960	1917	2557	0.75
1970	984	1799	0.55
1980	741	1835	0.40
1990	515	1191	0.43

Table: Ratio of 2016 Votes for Clinton to Trump by Birth Decade (Males)

Birth Decade	Total Clinton Votes	Total Trump Votes	Ratio
1930	361	184	1.96
1940	1478	986	1.50
1950	2020	1851	1.09
1960	1917	1731	1.11
1970	966	1262	0.77
1980	646	1279	0.51
1990	294	568	0.52

Table 2

Table: Ratio of 2020 Votes for Biden to Trump by Birth Decade (Females)

Birth Decade	Total Biden Votes	Total Trump Votes	Ratio
1930	239	277	0.86
1940	1389	1150	1.21
1950	2756	1686	1.63
1960	2592	1882	1.38
1970	1812	971	1.87
1980	1851	725	2.55
1990	1241	465	2.67

Table: Ratio of 2020 Votes for Biden to Trump by Birth Decade (Males)

Birth Decade	Total Biden Votes	Total Trump Votes	Ratio
1930	198	347	0.57
1940	1018	1446	0.70
1950	1910	1961	0.97
1960	1791	1857	0.96
1970	1301	927	1.40
1980	1316	609	2.16
1990	599	263	2.28

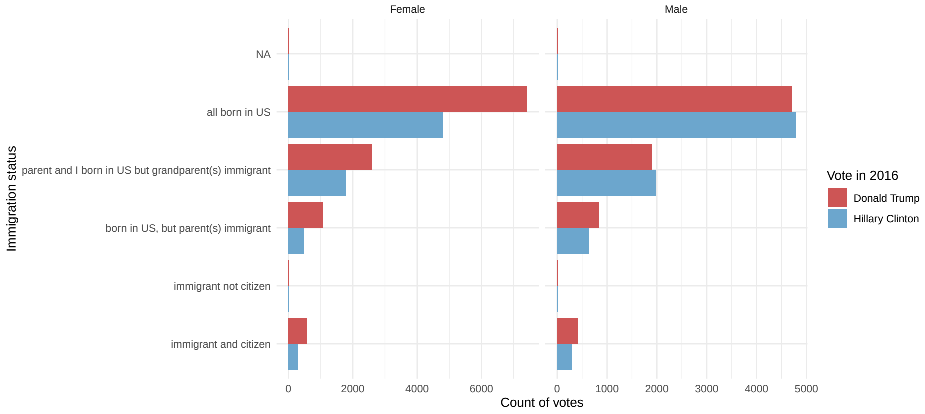
and in 2020 in Table 2. We divide respondents by birth decade, and we list the total number of votes for the Democratic Party and thee Republican Party, then list the ratio of the two. The decades 1920 and 2000 were omitted since the number of respondents in those Birth Decades were very low. The value of the ratio in centered at 1, where if the values is greater than 1, then within that group, the Democratic Party has more votes than the Republican Party. Where the values are less than 1, the Republican Party has more votes. When the ratio is around 1, the number of votes is approximately equal.

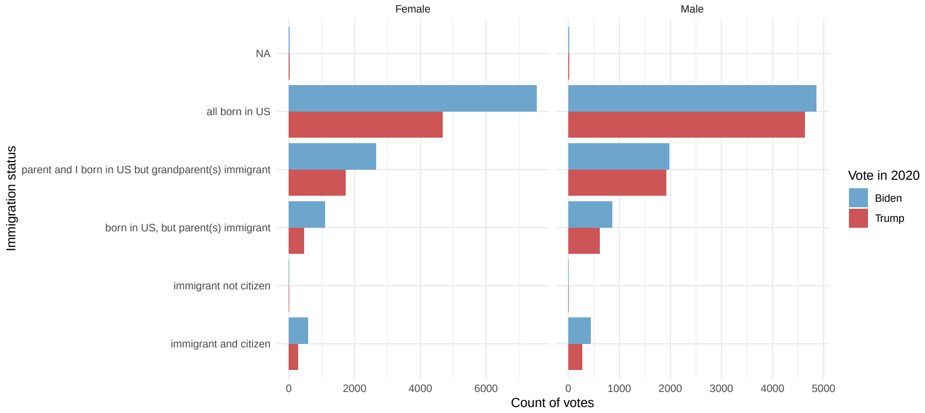
We notice that in 2016, the ratio of votes decreases as birth decade increases for both women and men voters. The opposite happens in 2020, where the ratio of votes increases as birth decade increases.

Model 3

We have modeled the following logistic regression in the graphs:

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{immigration status}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$





```
SAMPLING FOR MODEL 'bernoulli' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.000139 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 1.39 sec
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 0.509 seconds (Warm-up)
Chain 1:                0.608 seconds (Sampling)
Chain 1:                1.117 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'bernoulli' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 4.5e-05 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.45 sec
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2:
Chain 2:  Elapsed Time: 0.483 seconds (Warm-up)
Chain 2:                0.467 seconds (Sampling)
Chain 2:                0.95 seconds (Total)
Chain 2:

SAMPLING FOR MODEL 'bernoulli' NOW (CHAIN 3).
Chain 3:
Chain 3: Gradient evaluation took 4.2e-05 seconds
Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.42 sec
Chain 3: Adjust your expectations accordingly!
Chain 3:
Chain 3:
Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 3: 
Chain 3:  Elapsed Time: 0.497 seconds (Warm-up)
Chain 3:                0.454 seconds (Sampling)
Chain 3:                0.951 seconds (Total)
Chain 3: 

SAMPLING FOR MODEL 'bernoulli' NOW (CHAIN 4).
Chain 4: 
Chain 4: Gradient evaluation took 4.6e-05 seconds
Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.46 sec
Chain 4: Adjust your expectations accordingly!
Chain 4: 
Chain 4: 
Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4: 
Chain 4:  Elapsed Time: 0.416 seconds (Warm-up)
Chain 4:                0.5 seconds (Sampling)
Chain 4:                0.916 seconds (Total)
Chain 4: 
```

	Support Biden
(Intercept)	0.546
	(0.289)
genderMale	−0.422
	(0.128)
immstatborn in US, but parent(s) immigrant	0.457
	(0.352)
immstatparent and I born in US but grandparent(s) immigrant	−0.173
	(0.313)
immstatall born in US	−0.063
	(0.293)
Num.Obs.	997
R2	0.020
Log.Lik.	−670.734
ELPD	−675.8
ELPD s.e.	6.3
LOOIC	1351.5
LOOIC s.e.	12.7
WAIC	1351.5
RMSE	0.49

Results

Discussion

First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

Second discussion point

Third discussion point

Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

Additional data details

Model details

Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is
affected by, the data

Diagnostics

Checking the convergence of the
MCMC algorithm

References

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019a. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.