Case Study II - 2016 North Carolina Voter Registration

Emma Schmidt

STA 610 - Fall 2022

1 Introduction

The purpose of this study is to explore how voter demographics impacted the 2016 North Carolina presidential voting registration. This analysis is run using North Carolina State Board of Elections (NCSBE) data from the 2016 presidential election, combined with 2010 Census Bureau data reports for North Carolina. Useful covariates in the data include: county, age, race, ethnicity, gender, and party affiliation, as well as corresponding census and voter counts for each subgroup of individuals. Questions of specific interest focus on understanding how registration rates vary across subgroups, exploring how the probability odds of registering to vote differ county to county, and finally investigating the influence of party affiliation on registration rate amongst subgroups. The first steps in approaching these questions include pre-processing the data and conducting exploratory data analysis. Covariate selection and model construction are then described. Finally, I assess my models, noting key outputs, model fit, and overall takeaways.

2 Data Pre-Processing

The first major hurdle with the data are the systematic differences between the two data sets. For example, the election data contains sex data with three categories, (Male, Female, and Unspecified), whereas the census data only has Male and Female values. This unspecified category issue also occurs with the ethnic code covariate. Additionally, the party affiliation covariate only exists in the election data set. The second major hurdle occurs when evaluating total voter counts vs. total census counts for different counties. Specifically, some counties report more registered voters than the census reported number of residents. This is likely because the census data is six years older than the election data.

First addressing the unspecified category issue, I explore just how much unspecified data exists within each covariate. In the election data, voters with an unspecified sex make up roughly two percent of the voters. Assuming that the missingness in the gender data has occurred completely at random and because the degree of missingness is low, I omit all unspecified voters from the data. This strategy does not work with the ethnicity covariate because unspecified individuals represent approximately 20 percent of voters. Again, treating the unspecified data as missing completely at random, this time I impute the missing values by calculating respective percentages for Hispanic and Non-Hispanic voters, and then by assigning unspecified voters to be Hispanic or Non-Hispanic based on this probability distribution. Next, I address the lack of party affiliation in the census data. To

combat this issue I first omit libertarian voters from the data, as they only make up roughly .004 percent of total registered voters. Next, I calculate a party affiliation census count estimate by taking the total number of census respondents and subtracting election counts for all of the party affiliations not of interest. So, for example, if I want to estimate the total number of republicans, I take the total census count and subtract the total number of registered democrats and unaffiliated voters. This registration rate represents the total number of voters registered to a party out of all the individuals who could potentially be affiliated with that party. While this is not a perfect representation of true registration rate, it is a suitable estimate for modeling.

The issue of inflated voter registration counts caused by the time difference in data collection requires a less trivial approach. To resolve this problem I first explore the differences in voter registration counts and census resident counts grouped on county, age, race, ethnicity, and sex, flagging occurrences where voter registration count is greater than census population count. From there I filter the flagged occurrences out and build a Poisson regression model with the remaining data that predicts differences in voter registration and census population using the predictors: race, sex, age, ethnicity, log(total county size), and log(total voter count). Race, sex, age, and ethnicity are all factor variables. Race is categorized as white, black, or other; sex is categorized as male or female; age is categorized as 18-25, 26-40, 41-65, or 66+; and ethnicity is categorized as Hispanic or non-Hispanic. This model is specified below:

$$y_i : total\_population_i - total\_voters_i, \qquad y_i \sim Poisson(\lambda_i), \qquad y_i > 0$$

$$\log(\lambda_i) = \beta_0 + \beta_1 log(size_i) + \beta_2 log(vote_i) + \beta_r I(race_i = r) + \beta_s I(sex_i = s) + \beta_a I(age_i = a) + \beta_e I(ethnicity_i = e)$$

With this model, I generate predictions for all instances in the original data where voter registration exceeds census residency. Using the predictions, I impute new census count estimates by adding my predicted difference value to the voter registration count. Lastly, I merge the updated voter and census data, and to assist with the efficiency of modeling, I take a subset of 30 counties to proceed with my analysis.

3 EDA

Beginning the exploratory data analysis, I first seek to understand how each demographic subgroup registered to vote in the 2016 election. Specifically, I summarize party make up based on the sex, age, race, and ethnicity of its registered voters. Figure 1, seen below, offers a visual representation of these findings. Party registration by sex shows that democrats tend to have more registered females than males, roughly a 60-40 split. On the other hand, the republican and unaffiliated parties exhibit more of an even distribution with a near 50-50 split between male and female registered voters. The age distribution is similar across all three party affiliations, with most of the registered voters being in the 41-65 age range, followed by the 26-40 age range. For both democrats and republicans the age range with the lowest percentage of registered voters is 18-25, however for unaffiliated voters this range is 66+. This suggests that unaffiliated voters tend to be younger than democrat and republican

voters. Looking into the race demographic, I noticed that of the seven race categories, black and white voters make up about 91 percent of the data. Because the other five race categories represent very small proportions of the data, I combined them into an other category. The distribution shows that democrats are about evenly white and black voters, whereas republicans and unaffiliated voters are primarily white. In examining ethnicity in all three parties non-Hispanic voters make up over 95 percent of the registered voters. This is not surprising, as Hispanic voters represent only about three percent of registered voters.

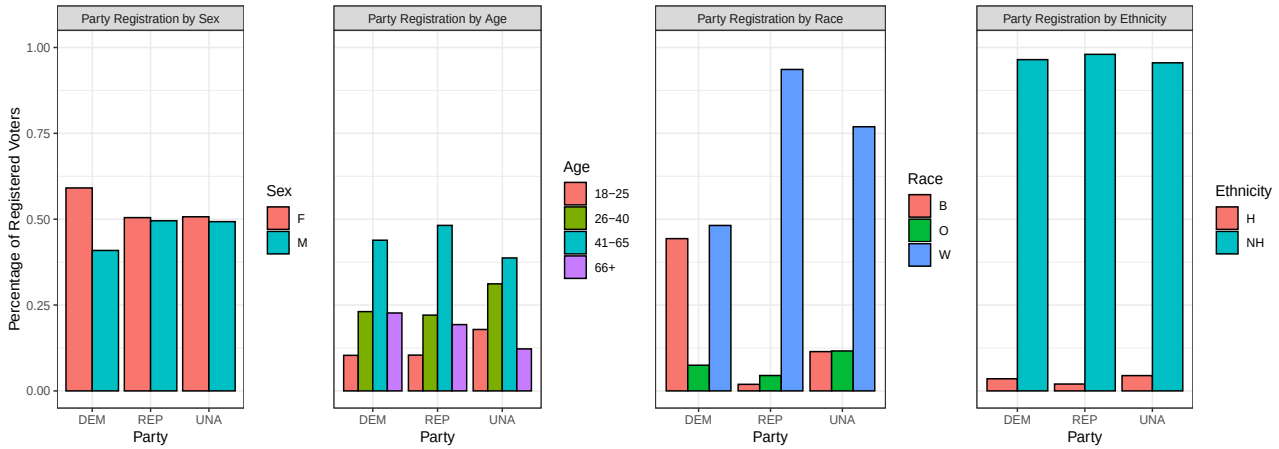


Figure 1: Party Registration by Demographic

Next I examine how registration rates varied from county to county in 2016. Figure 2 offers the registration rates for each of the 30 randomly sampled counties in the data set. Duplin, Hyde, and Onslow county stand out as counties with overwhelmingly low voter registration rates. These counties have registrations rates all below 50 percent. Only Iredell and Madison county possess registration rates above 75 percent, with rates of 75.5 percent and 76.8 percent respectively. Further investigation reveals that there are no clear discrepancies between the counties with low registration rates vs. counties with high registration rates. Both groups contained counties with high and low populations, and broad mixtures of demographics.

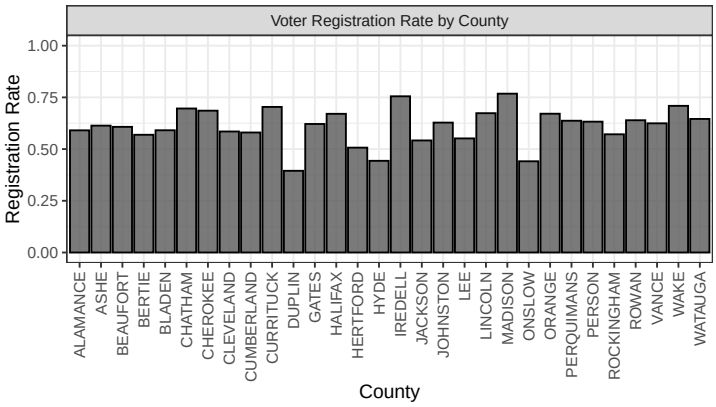


Figure 2: Registration Rate by County

The final question of interest is centered around how registration rates vary across different demographics. Using the registration rate discussed in section two of this paper, Figure 3 highlights these rates for the sex, age, race, and ethnicity demographics. The first key finding is that across all three party affiliations, females demonstrate higher registration rates than males, with democratic females having the highest rate. Next observing age, all three parties show a positive correlation between age and registration rate, with democrats having higher rates than the other two categories. Moving to race, black democrats have the highest registration rate amongst all races and parties, with a rate of 88 percent. In the republican party white voters have over two times the registration rate of black and other republicans. Lastly, ethnicity shows that the registration rate for non-Hispanics is overwhelmingly higher than that or Hispanics for all three parties.

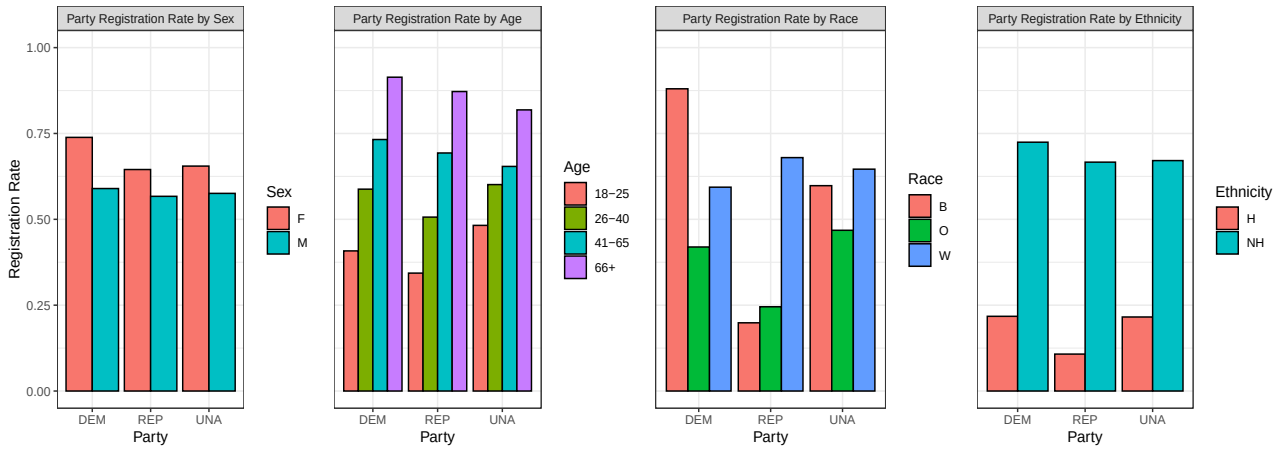


Figure 3: Party Registration Rate by Demographic

Overall, sex, age, and race all look like interesting covariates to potential models. Ethnicity is also an interesting covariate, but may not be as informative due to the high ratio of non-Hispanic to Hispanic voters.

4 Model Selection

The first model is a binomial regression model with a logit link function that seeks to answer questions surrounding voter registration rates for different demographics with a random intercept on county. With no background knowledge on voter registration rates, I specify weakly informative priors. Model 1 can be specified as:

$$logit(p_{ij}) = b_{county_j} + \beta_1 I(Male_i) + \beta_2 I(White_i) + \beta_3 I(Other_i) + \beta_4 I(Rep_i) + \beta_5 I(Una_i) \\ + \beta_6 I(26-40_i) + \beta_7 I(41-65_i) + \beta_8 I(66+_i) + \beta_9 I(NH_i)$$

$$\mathrm{P}(\mathrm{Y}_{ij} = y_{ij}|p_{ij}) \sim Binomial(n_{ij}, p_{ij}), \quad \beta_{all} : representative\ of\ all\ specified\ \beta's$$

$$\mathrm{b}_{county_j} \sim N(b, \sigma), \quad b \sim N(0, 10), \quad \sigma \sim HalfCauchy(0, 1), \quad \beta_{all} \sim N(0, 1)$$

Figure 4 can be referenced below to explore Model 1's summary output. The first notable call out is that older, non-Hispanic voters have the highest log odds ratios for registration turn out. These findings are consistent with the findings from the exploratory data analysis on registration rates. The model suggests that male voters have higher registration rates than females, but this is deceiving because the intercept is also absorbing younger voters, who tend to register at much lower rates. White voters have higher odds ratios than all of the other races, but this is not surprising because white voters represent over 70 percent of the data. Looking at party affiliation, republican and unaffiliated voters have comparable log ratios that suggest higher voter registration than that of democrats. This makes sense because the republican and unaffiliated parties are predominantly white, and as mentioned before, white voters have high registration rates and make up a large proportion of the data.

	Registration Rate	
Predictors	*Odds Ratios*	*CI (95%)*
Intercept	0.21	0.17 – 0.25
sexMale	0.67	0.67 – 0.67
raceOther	0.42	0.42 – 0.43
raceWhite	0.54	0.53 – 0.54
party_cdREP	0.87	0.87 – 0.88
party_cdUNA	1.01	1.00 – 1.02
age26M40	2.03	2.02 – 2.05
age41M65	2.96	2.94 – 2.98
age66P	9.19	9.09 – 9.29
ethnicityNonMHispanic	7.95	7.87 – 8.02
Random Effects		
σ^2	3.29	
$\tau_{00\ county\_desc}$	0.23	
ICC	0.07	
$N_{\ county\_desc}$	30	
Observations	3870	
Marginal R^2 / Conditional R^2	0.934 / 0.981	

Figure 4: Voter Registration Rate Model Summary

Moving to county specific registration rates, Figure 5 uses posterior samples to display box plots of the log odds ratios of voter registration for each of the selected counties. Consistent with the exploratory data analysis, Iredell and Madison were the top performing counties, and Duplin and Hyde were the worst performing. As mentioned before, it is unclear in the data why these particular counties do so successfully and poorly, as they don't exemplify any clear differences in demographic make up.

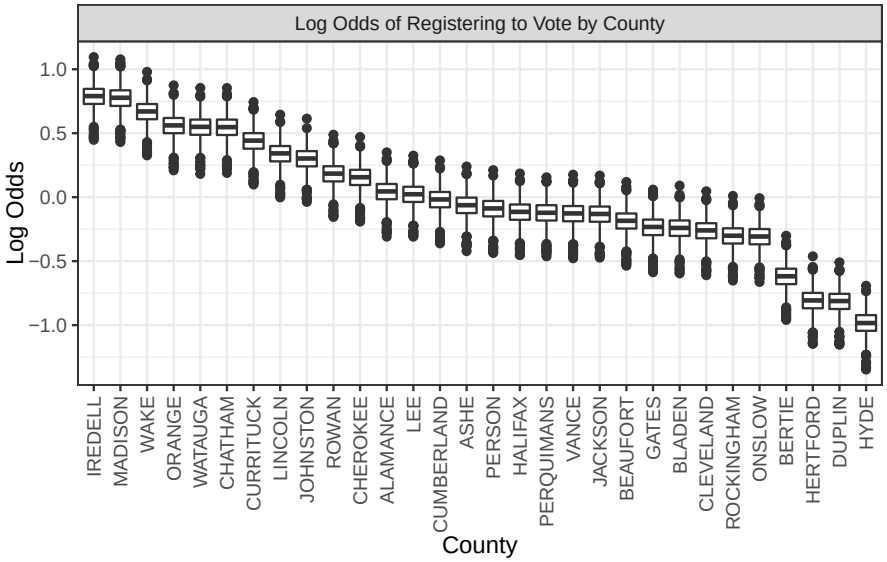


Figure 5: Box plots by County

The second model seeks to explore how the registration rates for different demographic groups vary within party affiliation. To achieve this, the second model is run three separate times, with republican specific, democrat specific, and unaffiliated specific voter data. Again, with no background knowledge on voter registration rates, I specify weakly informative priors. Model 2 is specified below:

$$logit(p_{ij}) = b_{county_j} + \beta_1 I(Male_i) + \beta_2 I(White_i) + \beta_3 I(Other_i) + \beta_4 I(26 - 40_i) + \beta_5 I(41 - 65_i) + \beta_6 I(66+_i) + \beta_7 I(NH_i)$$

$$\mathrm{P}(\mathrm{Y}_{ij} = y_{ij}|p_{ij}) \sim Binomial(n_{ij}, p_{ij}), \quad \beta_{all} : representative\ of\ all\ specified\ \beta's$$

$$\mathrm{b}_{county_j} \sim N(b, \sigma), \quad b \sim N(0, 10), \quad \sigma \sim HalfCauchy(0, 1), \quad \beta_{all} \sim N(0, 1)$$

Figure 6 shows the model summaries for each of the three models. For republicans the demographic groups that have overwhelmingly high registration rates compared to those of the others, are white, non-Hispanic voters, over the age of 66. On the other side of the spectrum, the intercept, representing black, female voters, between the ages of 18-24, has an extremely low odds ratio of 0.01. These model conclusions are consistent with the findings in the data. Moving to the democrat model, the demographic groups with the two highest odds ratios are non-Hispanic voters and voters over the age of 66, and the demographic groups with the two lowest odds ratios are white and other race voters. This makes sense, because black democrats showed high registration rates in the data. In the unaffiliated model, again non-Hispanic voters and voters aged over 66 had the highest odds ratios. The ratios for these demographics were high in all three models, suggesting that these voters have high turnout rates regardless of party affiliation. Much like the republican model, the intercept in the unaffiliated model has the lowest odds ratio of all the covariates at 0.14. With similar demographic make up and registration rate distributions, it makes sense that the republican and unaffiliated models behave similarly.

	Republican Registration Rate		Democrat Registration Rate		Unaffiliated Registration Rate	
Predictors	*Odds Ratios*	*CI (95%)*	*Odds Ratios*	*CI (95%)*	*Odds Ratios*	*CI (95%)*
Intercept	0.01	0.01 – 0.01	0.58	0.49 – 0.69	0.14	0.11 – 0.17
sexMale	0.77	0.76 – 0.78	0.54	0.54 – 0.55	0.72	0.72 – 0.73
raceOther	2.71	2.64 – 2.78	0.19	0.19 – 0.19	1.00	0.99 – 1.02
raceWhite	7.51	7.34 – 7.68	0.16	0.15 – 0.16	1.05	1.04 – 1.07
age26M40	2.20	2.17 – 2.23	2.24	2.21 – 2.27	1.74	1.72 – 1.76
age41M65	3.82	3.77 – 3.87	3.86	3.81 – 3.91	1.87	1.85 – 1.89
age66P	10.04	9.85 – 10.24	17.52	17.18 – 17.86	4.27	4.19 – 4.35
ethnicityNonMHispanic	12.17	11.93 – 12.41	6.21	6.10 – 6.31	7.66	7.54 – 7.78
Random Effects						
σ^2	3.29		3.29		3.29	
τ_{00}	0.30 $_{\text{county\_desc}}$		0.22 $_{\text{county\_desc}}$		0.31 $_{\text{county\_desc}}$	
ICC	0.08		0.06		0.09	
N	30 $_{\text{county\_desc}}$		30 $_{\text{county\_desc}}$		30 $_{\text{county\_desc}}$	
Observations	1175		1377		1318	
Marginal R^2 / Conditional R^2	0.962 / 0.992		0.968 / 0.993		0.918 / 0.985	

Figure 6: Party Registration Rate Model Summary

5 Model Fit

With Model 1 and Model 2 specified above, I now assess model fit via posterior predictive checks. First looking at Model 1, Figure 7 offers two plots for assessing model fit. The first is a plot of fitted vs. residual values. The plot shows that the points are roughly scattered around zero. Additionally, looking at a histogram of the log(predicted voters) vs. the log(true voters), it appears that the model slightly underestimates total voters.

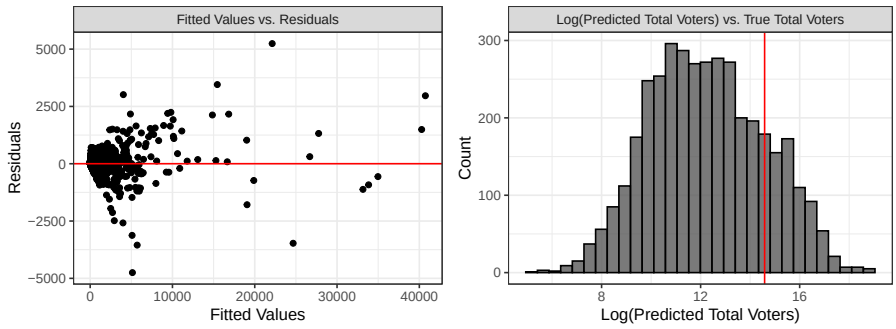


Figure 7: Model 1 Fit

To begin assessing Model 2 for each party affiliation data set, Figure 8 displays fitted values vs. residuals plots, as well as histograms of predicted log(predicted total voters) vs. log(true total voters) for each respective party. In all three cases the fitted vs. residuals plots look to have points centered around zero, with no clear signs of any issues in the data. Pivoting to the histograms, the model has marginally underestimated true voters for all three parties. Additionally, the democrat histogram is bimodal, predicting two potential peaks in total voter count. This bimodal distribution is consistent with distribution of true voter counts from the democrat data; this plot can be found in the appendix.

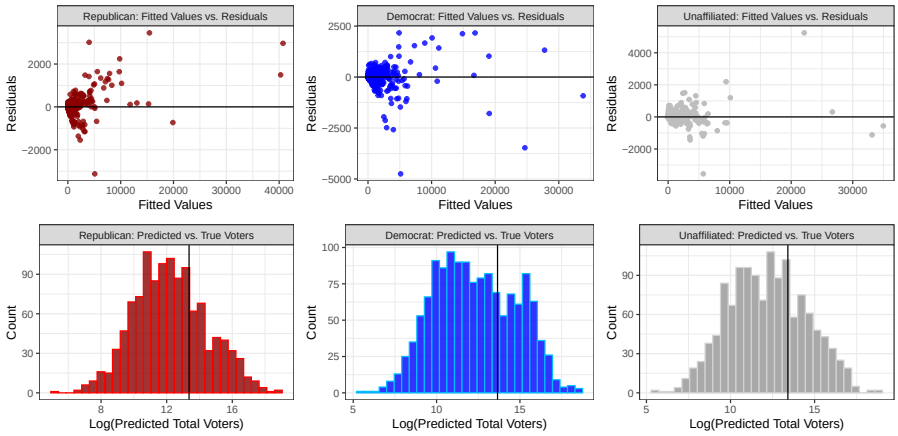


Figure 8: Model 2 Fit

6 Conclusion

The final models indicate that voter demographics are reflective of potential registration rates. Specifically, as a whole, registration rates tended to be highest for older, white, non-Hispanic voters that were registered republican or unaffiliated. Partitioning by party affiliation, the second model indicates that older, white, non-Hispanic males had the highest registration rates amongst republican and unaffiliated voters. For democrats the model shows that older, black, non-Hispanic females had the highest registration rates. Overall, white males exemplified the highest registration rates in the republican and unaffiliated parties, black females had the highest registration rates amongst democrats, and older non-Hispanic voters had high registration rates regardless of party affiliation.

Limitations to the exploration of registration rates across and within demographic subgroups and to the modeling process are mostly brought on by the data. The six year gap between data collection for the respective data sets raises reliability concerns before modeling has even begun. Additionally, the lack of party affiliation in the census data, limits the interpretability of the results. Specifically, registration rate by party affiliation cannot be interpreted as, for example, the total number of voting democrats out of all democrats, rather as the total number of voting democrats out of people who are potentially democrat. Nonetheless, the final models yield interesting results worth exploring.

Appendix

Emma Schmidt

2022-11-06

```
census <- read.table("~/Desktop/Duke/610/Census2010_long.txt", header = TRUE)

voter <- read.table("~/Desktop/Duke/610/voter_stats_20161108.txt", header = TRUE)
```

```
# Explore Voter Categories

# Race
voter %>%
  filter(race_code != "") %>%
  mutate(total_sum = sum(total_voters)) %>%
  group_by(race_code) %>%
  summarize(race_sum = 100*(sum(total_voters)/mean(total_sum)))
```

```
## # A tibble: 7 x 2
##   race_code race_sum
##   <chr>        <dbl>
## 1 A            1.17
## 2 B           22.2
## 3 I            0.818
## 4 M            0.687
## 5 O            2.40
## 6 U            3.37
## 7 W           69.4
```

```
# Ethnic Code
voter %>%
  mutate(total_sum = sum(total_voters)) %>%
  group_by(ethnic_code) %>%
  summarize(ethnicity_sum = sum(total_voters)/mean(total_sum))
```

```
## # A tibble: 3 x 2
##   ethnic_code ethnicity_sum
##   <chr>              <dbl>
## 1 HL                0.0241
## 2 NL                0.779
## 3 UN                0.197
```

```
# Gender
voter %>%
  mutate(total_sum = sum(total_voters)) %>%
  group_by(sex_code) %>%
  summarize(sex_sum = sum(total_voters)/mean(total_sum))
```

```
## # A tibble: 3 x 2
##   sex_code sex_sum
##   <chr>      <dbl>
## 1 F          0.530
## 2 M          0.448
## 3 U          0.0229
```

```
# Party Code
voter %>%
  mutate(total_sum = sum(total_voters)) %>%
  group_by(party_cd) %>%
  summarize(party_sum = sum(total_voters)/mean(total_sum))
```

```
## # A tibble: 4 x 2
##   party_cd party_sum
##   <chr>        <dbl>
## 1 DEM        0.395
## 2 LIB        0.00468
## 3 REP        0.302
## 4 UNA        0.299
```

```
# Age
voter %>%
  mutate(total_sum = sum(total_voters)) %>%
  group_by(age) %>%
  summarize(party_sum = sum(total_voters)/mean(total_sum))
```

```
## # A tibble: 4 x 2
##   age          party_sum
##   <chr>            <dbl>
## 1 Age 18 - 25      0.125
## 2 Age 26 - 40      0.250
## 3 Age 41 - 65      0.429
## 4 Age Over 66      0.196
```

```
# Data Pre-processing
voter_update <- voter %>%
  mutate(race = case_when(race_code == "W" ~ "White",
                          race_code == "B" ~ "Black",
                          TRUE ~ "Other"),
         age = case_when(age == "Age 41 - 65" ~ "41-65",
                         age == "Age 18 - 25" ~ "18-25",
                         age == "Age 26 - 40" ~ "26-40",
                         TRUE ~ "66+"),
         sex = case_when(sex_code == "M" ~ "Male",
                         sex_code == "F" ~ "Female",
                         TRUE ~ "Unspecified"),
         ethnicity = case_when(ethnic_code == "HL" ~ "Hispanic",
                          ethnic_code == "NL" ~ "Non-Hispanic",
                          TRUE ~ "Unspecified")) %>%
  filter(sex != "Unspecified", party_cd != "LIB")
```

```r
# Impute Hispanic/Non-Hispanic
gg <- voter_update %>%
  filter(ethnicity != "Unspecified") %>%
  group_by(ethnicity) %>%
  summarise(total_voters = sum(total_voters)) %>%
  pull(total_voters)

perc_hispanic <- gg[1]/sum(gg)



add_hispanic <- voter_update %>%
  filter(ethnicity == "Unspecified") %>%
  group_by(county_desc, race, age, sex, party_cd) %>%
  summarise(total_voters = round(perc_hispanic * sum(total_voters))) %>%
  mutate(ethnicity = "Hispanic") %>%
  filter(total_voters > 0)
```

```
## `summarise()` has grouped output by 'county_desc', 'race', 'age', 'sex'. You
## can override using the `.groups` argument.
```

```r
add_nothispanic <- voter_update %>%
  filter(ethnicity == "Unspecified") %>%
  group_by(county_desc, race, age, sex, party_cd) %>%
  summarise(total_voters = round((1-perc_hispanic) * sum(total_voters))) %>%
  mutate(ethnicity = "Non-Hispanic") %>%
  filter(total_voters > 0)
```

```
## `summarise()` has grouped output by 'county_desc', 'race', 'age', 'sex'. You
## can override using the `.groups` argument.
```

```r
new_hispanic <- rbind(add_hispanic, add_nothispanic) %>%
  select(county_desc, race, age, sex, party_cd, ethnicity, total_voters)



old_hispanic <- voter_update %>%
  filter(ethnicity != "Unspecified") %>%
  group_by(county_desc, race, age, sex, party_cd, ethnicity) %>%
  summarise(total_voters = sum(total_voters))
```

```
## `summarise()` has grouped output by 'county_desc', 'race', 'age', 'sex',
## 'party_cd'. You can override using the `.groups` argument.
```

```r
voter_agg <- rbind(old_hispanic, new_hispanic) %>%
  group_by(county_desc, race, age, sex, party_cd, ethnicity) %>%
  summarise(total_voters = sum(total_voters))
```

```
## `summarise()` has grouped output by 'county_desc', 'race', 'age', 'sex',
## 'party_cd'. You can override using the `.groups` argument.
```

```
census_agg <- census %>%
  mutate(county_desc = Geography,
         age = Age,
         sex = Gender,
         race = case_when(Race == "WhiteAlone" ~ "White",
                          Race == "BlackAlone" ~ "Black",
                          TRUE ~ "Other"),
         ethnicity = case_when(Hispanic == "NotHispanic" ~ "Non-Hispanic",
                           TRUE ~ "Hispanic")) %>%
  group_by(county_desc, sex, race, age, ethnicity) %>%
  summarise(Freq = sum(Freq), TotalCountyPopulation = mean(TotalCountyPopulation))
```

```
## `summarise()` has grouped output by 'county_desc', 'sex', 'race', 'age'. You
## can override using the `.groups` argument.
```

```
# Merge the two data frames
data_merge <- merge(voter_agg, census_agg, by=c('county_desc', 'age',
                                                'sex', 'race', 'ethnicity'))
# Impute census population data
check <- data_merge %>%
  group_by(county_desc, race, age, sex, ethnicity) %>%
  summarise(total_voters = sum(total_voters),
            Freq = mean(Freq),
            log_tot_voters = mean(log(total_voters)),
            log_tot_county = mean(log(TotalCountyPopulation))) %>%
  mutate(is_over = (total_voters > Freq))
```

```
## `summarise()` has grouped output by 'county_desc', 'race', 'age', 'sex'. You
## can override using the `.groups` argument.
```

```
sum(check$is_over)
```

```
## [1] 1711
```

```
check <- check %>%
  mutate(diff = Freq - total_voters)

# Poisson regression to impute the total number of voters when the difference is positive
fit <- glm(diff ~ race + sex + age + ethnicity + log_tot_county + log_tot_voters,
           data = check %>% filter(is_over == FALSE), family = "poisson")

# Make the prediction where the diff is negative
pred_diff <- round(predict(fit, newdata =  check %>% filter(is_over == TRUE),
                     type = "response"))

# Create a new column with the frequency corrected by the prediction
new_diff <- check$diff
new_diff[check$is_over==TRUE] <- pred_diff
check$Freq2 <- new_diff + check$total_voters

# Join with the whole dataset
```

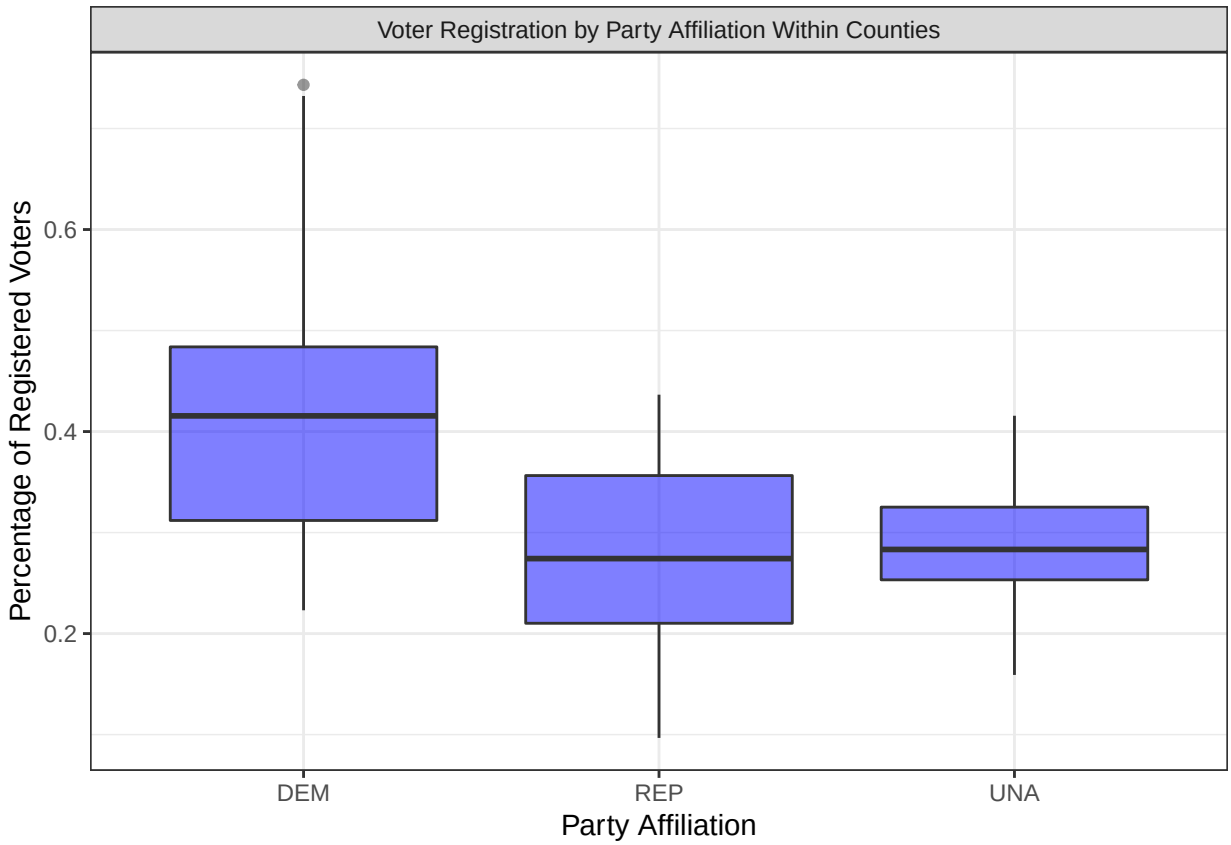
```r
data_merge2 <- merge(data_merge, check %>%
              select(county_desc, race, age, sex, ethnicity, Freq2),
              by = c("county_desc", "race", "age", "sex", "ethnicity"))

# Compute registration turnout
df_reg <- data_merge2 %>%
  group_by(county_desc, race, age, sex, ethnicity) %>%
  mutate(Freq = Freq2 - sum(total_voters) + total_voters)
```

```r
# select 30 counties
set.seed(45)

counties <- unique(voter$county_desc)

selected_counties <- sample(counties, 30, replace = FALSE)

df_select <- df_reg %>%
  filter(county_desc %in% selected_counties)
```

```r
# percentage of voters by party
pct <- df_select %>%
  group_by(county_desc, party_cd) %>%
  summarize(total = sum(total_voters)) %>%
  ungroup() %>%
  group_by(county_desc) %>%
  mutate(percentage = total/sum(total))
```

```
## `summarise()` has grouped output by 'county_desc'. You can override using the
## `.groups` argument.
```

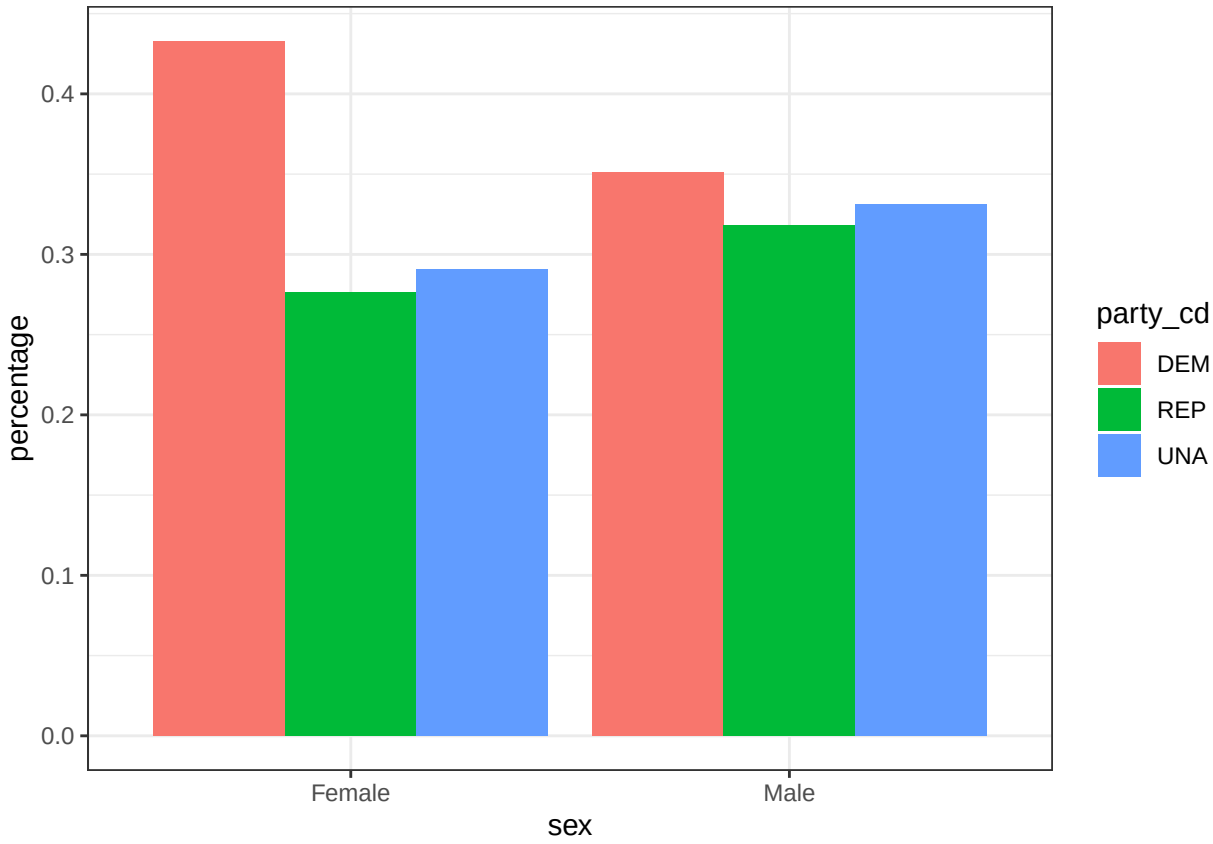
```r
ggplot(data = pct, aes(x=party_cd, y=percentage)) +
  geom_boxplot(fill = "blue", alpha = 0.5) +
  theme_bw() +
  labs(x = "Party Affiliation", y = "Percentage of Registered Voters") +
  facet_wrap(~"Voter Registration by Party Affiliation Within Counties")
```




```
# Explore Sex Code
sex <- df_select %>%
  group_by(sex, party_cd) %>%
  summarize(total_voters = sum(total_voters)) %>%
  ungroup() %>%
  group_by(sex) %>%
  mutate(percentage = total_voters/sum(total_voters))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

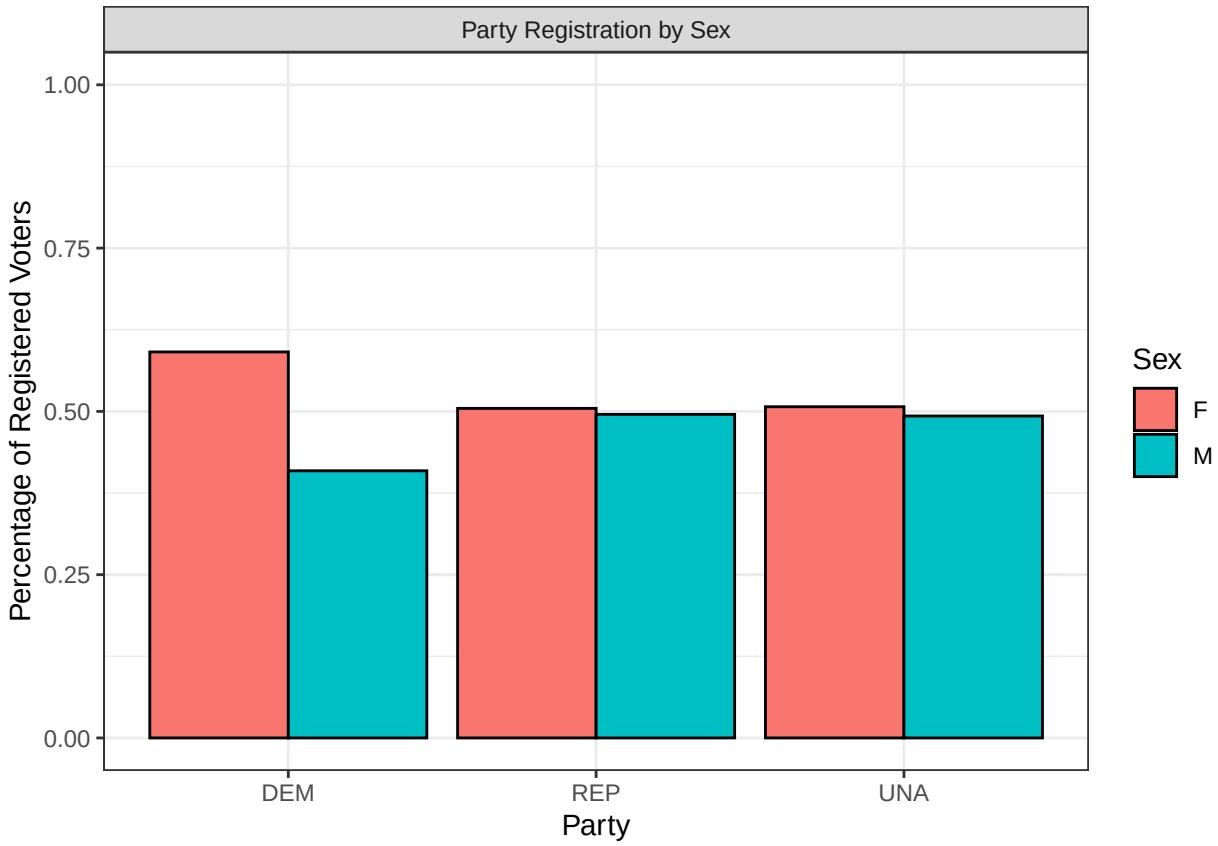
```
ggplot(data = sex, aes(x=sex, y=percentage, fill = party_cd)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_bw()
```




```
sex2 <- df_select %>%
  group_by(sex, party_cd) %>%
  summarize(total_voters = sum(total_voters)) %>%
  ungroup() %>%
  group_by(party_cd) %>%
  mutate(percentage = total_voters/sum(total_voters))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

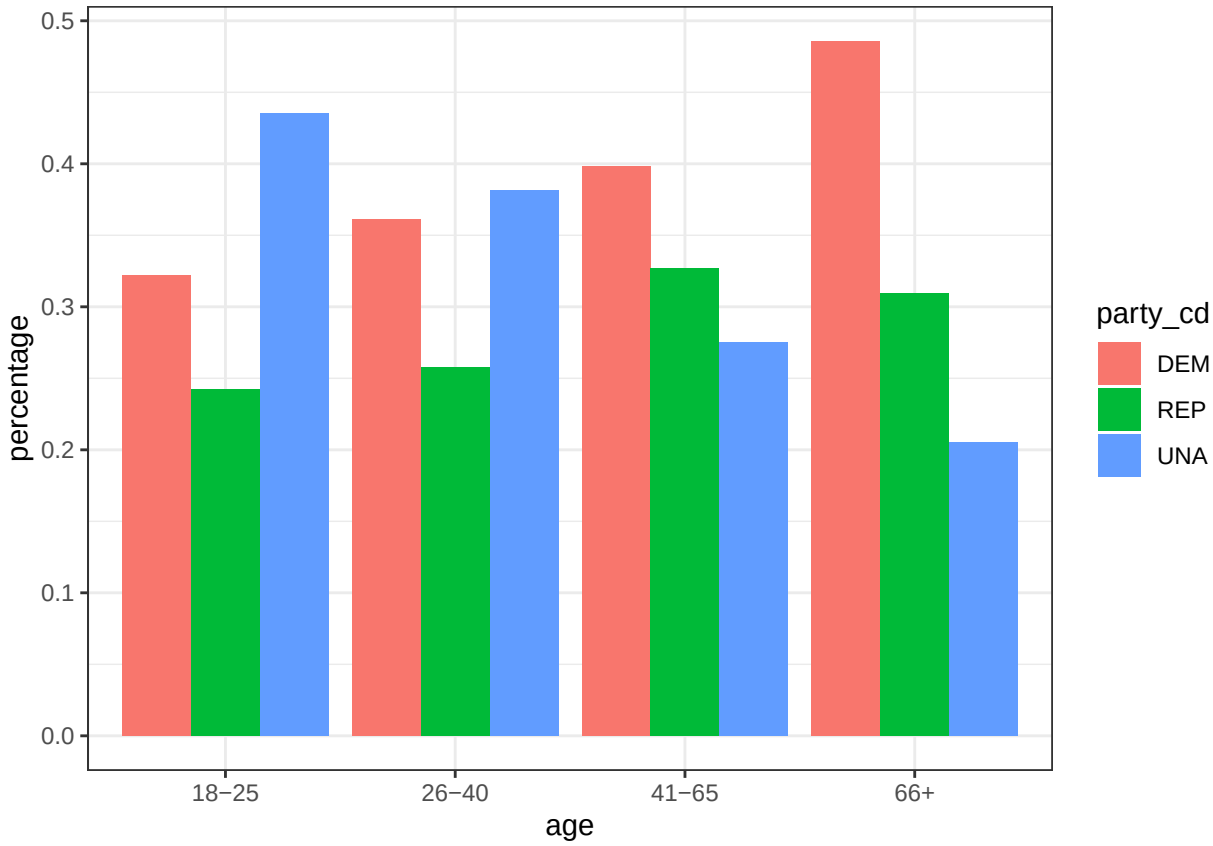
```
sex_plot <- ggplot(data = sex2, aes(x=party_cd, y=percentage, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  labs(x="Party", y="Percentage of Registered Voters", fill="Sex") +
  scale_fill_discrete(labels=c('F', 'M')) +
  facet_wrap(~"Party Registration by Sex")

sex_plot
```




```
# percentage of voters by age group
age <- df_select %>%
  group_by(age, party_cd) %>%
  summarize(total = sum(total_voters)) %>%
  ungroup() %>%
  group_by(age) %>%
  mutate(percentage = total/sum(total))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

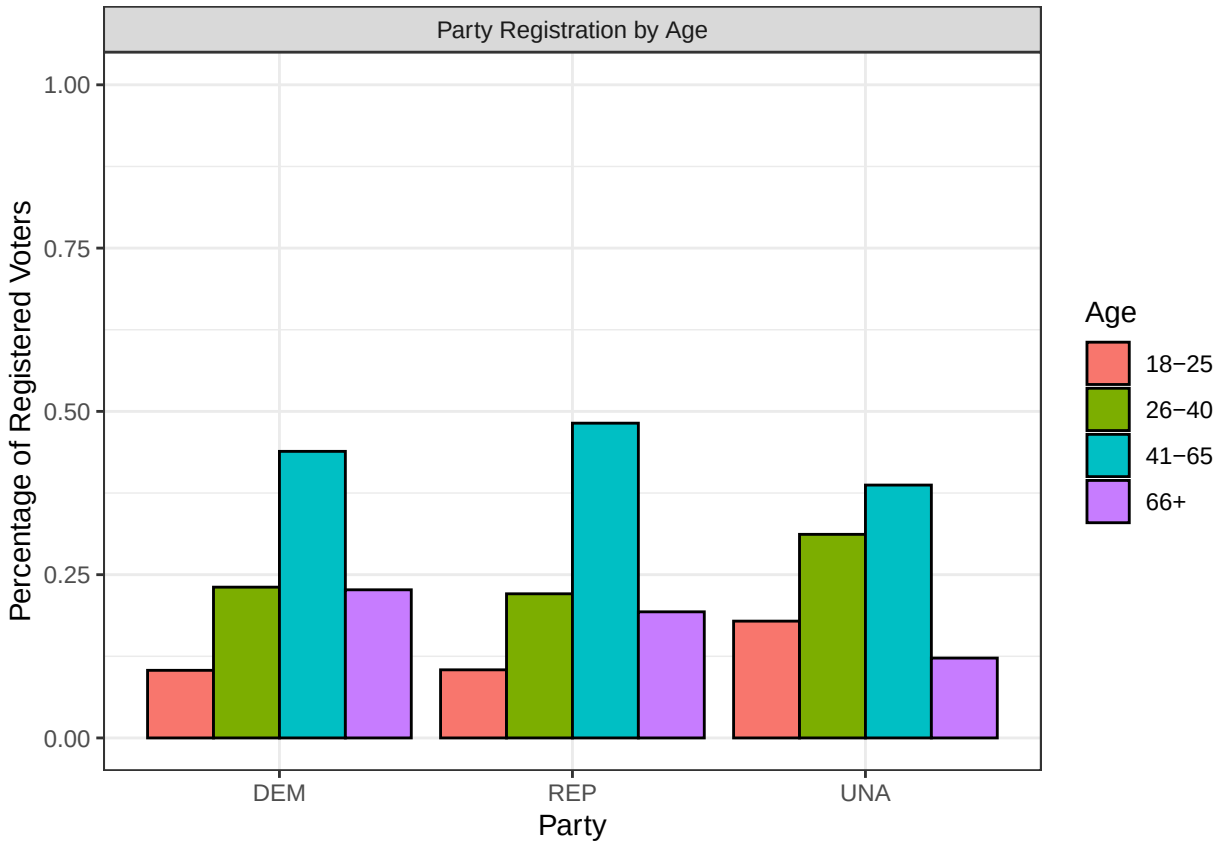
```
ggplot(data = age, aes(x=age, y=percentage, fill = party_cd)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_bw()
```




```
age2 <- df_select %>%
  group_by(age, party_cd) %>%
  summarize(total = sum(total_voters)) %>%
  ungroup() %>%
  group_by(party_cd) %>%
  mutate(percentage = total/sum(total))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

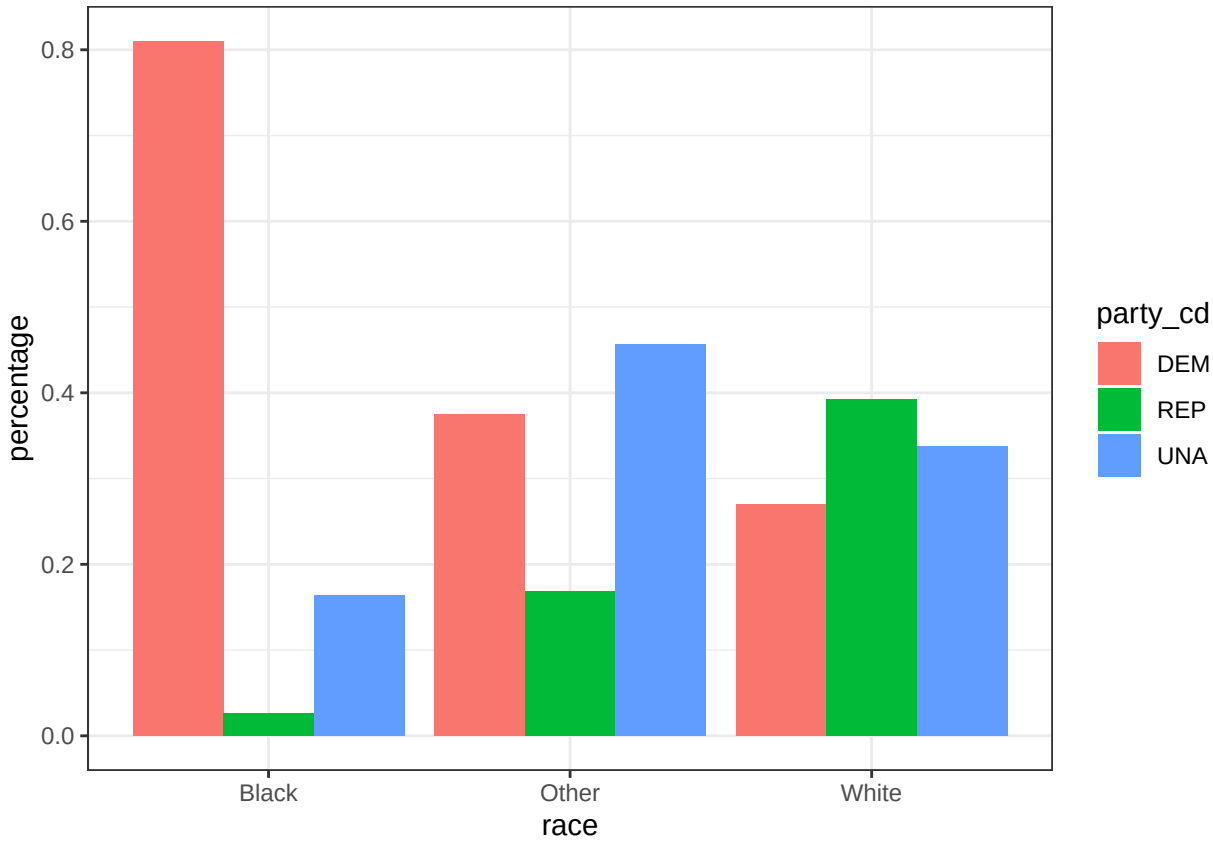
```
age_plot <- ggplot(data = age2, aes(x=party_cd, y=percentage, fill = age)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  labs(x="Party", y="Percentage of Registered Voters", fill = "Age") +
  facet_wrap(~"Party Registration by Age")

age_plot
```




```
# Explore Race Code
race <- df_select %>%
  group_by(race, party_cd) %>%
  summarize(total_voters = sum(total_voters)) %>%
  ungroup() %>%
  group_by(race) %>%
  mutate(percentage = total_voters/sum(total_voters))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

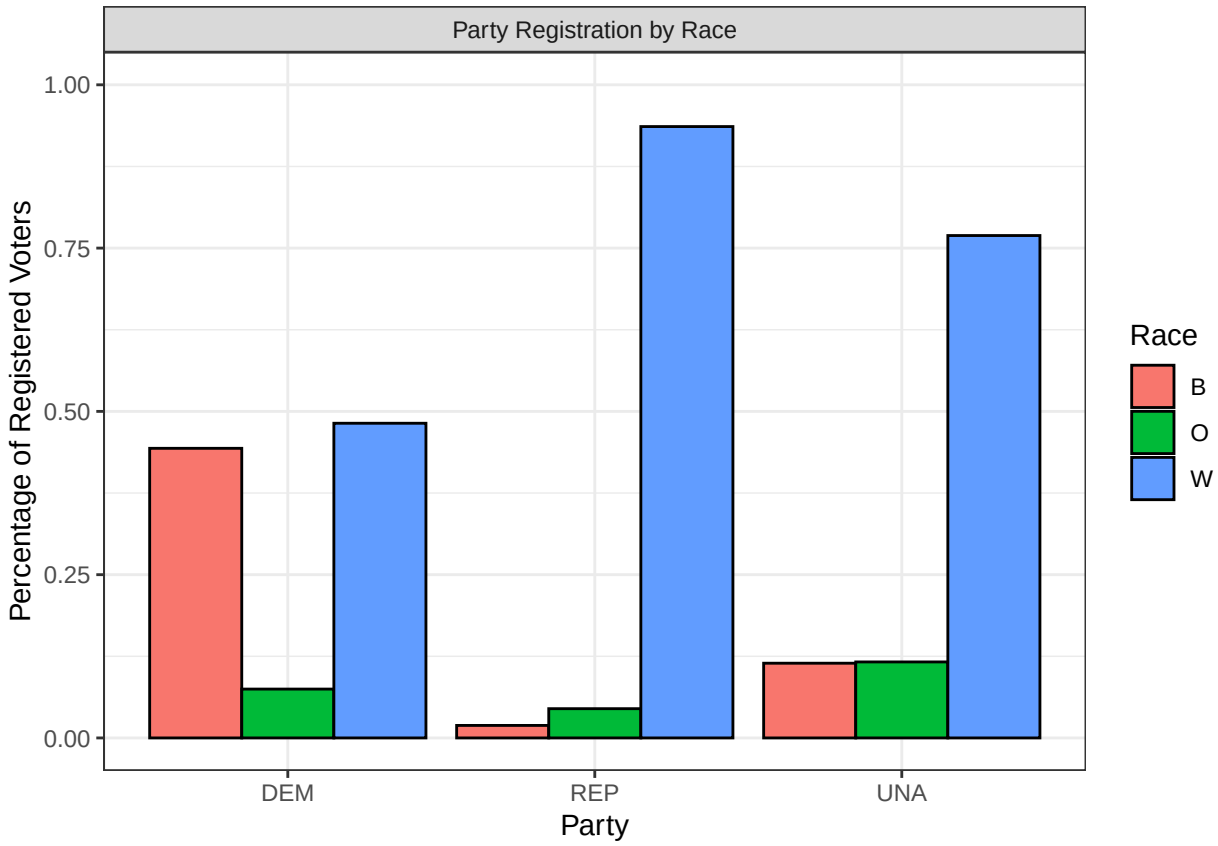
```
ggplot(data = race, aes(x=race, y=percentage, fill = party_cd)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_bw()
```




```
race2 <- df_select %>%
  group_by(race, party_cd) %>%
  summarize(total_voters = sum(total_voters)) %>%
  ungroup() %>%
  group_by(party_cd) %>%
  mutate(percentage = total_voters/sum(total_voters))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

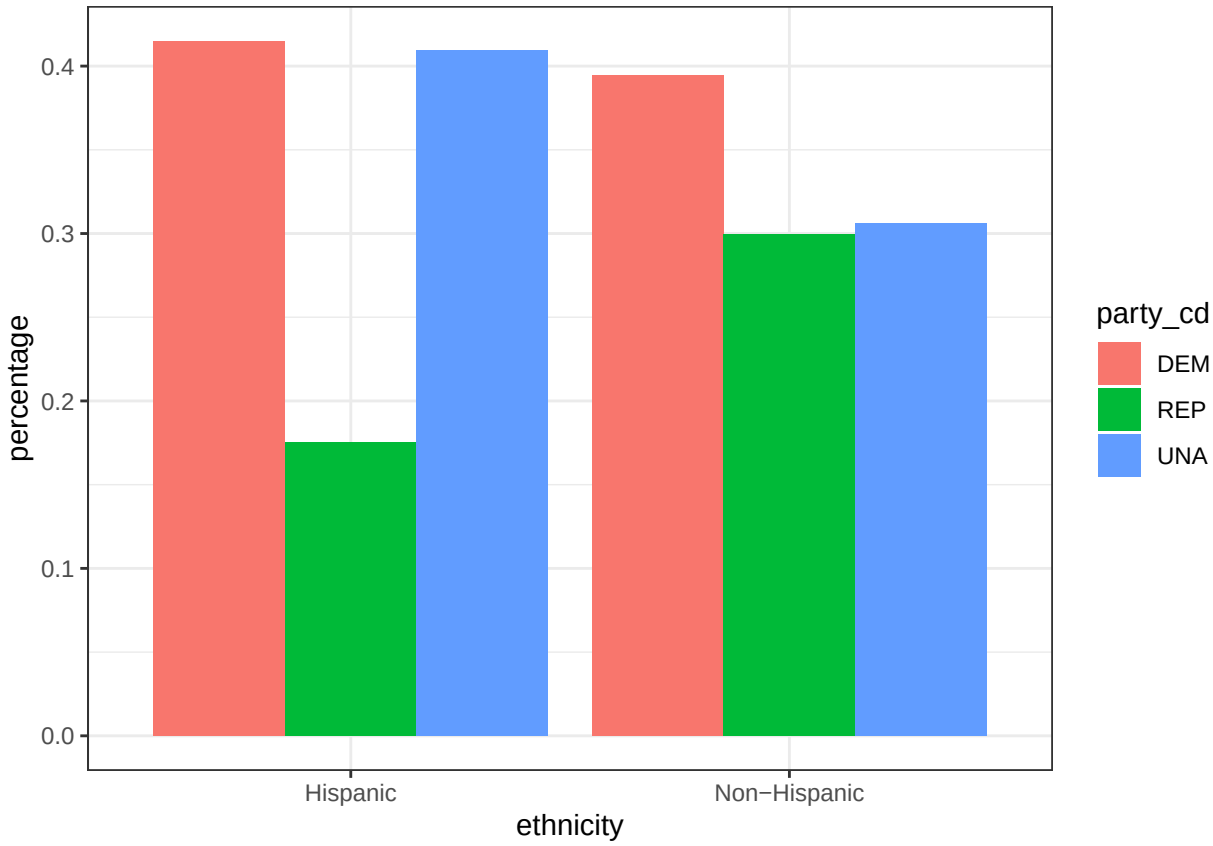
```
race_plot <- ggplot(data = race2, aes(x=party_cd, y=percentage, fill = race)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  labs(x="Party", y="Percentage of Registered Voters", fill = "Race") +
  scale_fill_discrete(labels=c("B", "O", "W")) +
  facet_wrap(~"Party Registration by Race")

race_plot
```




```
# Explore Ethnic Code
hisp <- df_select %>%
  group_by(ethnicity, party_cd) %>%
  summarise(total_voters = sum(total_voters)) %>%
  ungroup() %>%
  group_by(ethnicity) %>%
  mutate(percentage = total_voters/sum(total_voters))
```

```
## `summarise()` has grouped output by 'ethnicity'. You can override using the
## `.groups` argument.
```

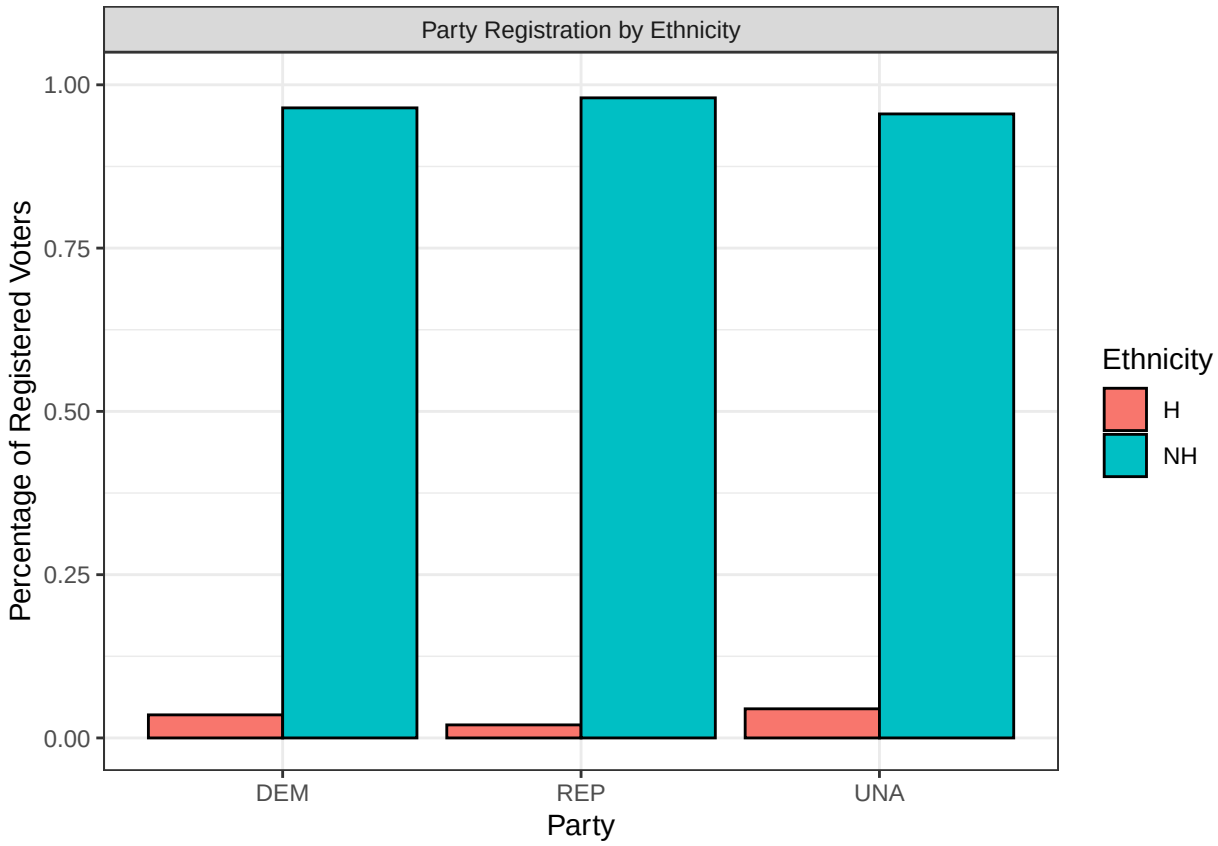
```
ggplot(data = hisp, aes(x=ethnicity, y=percentage, fill = party_cd)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_bw()
```



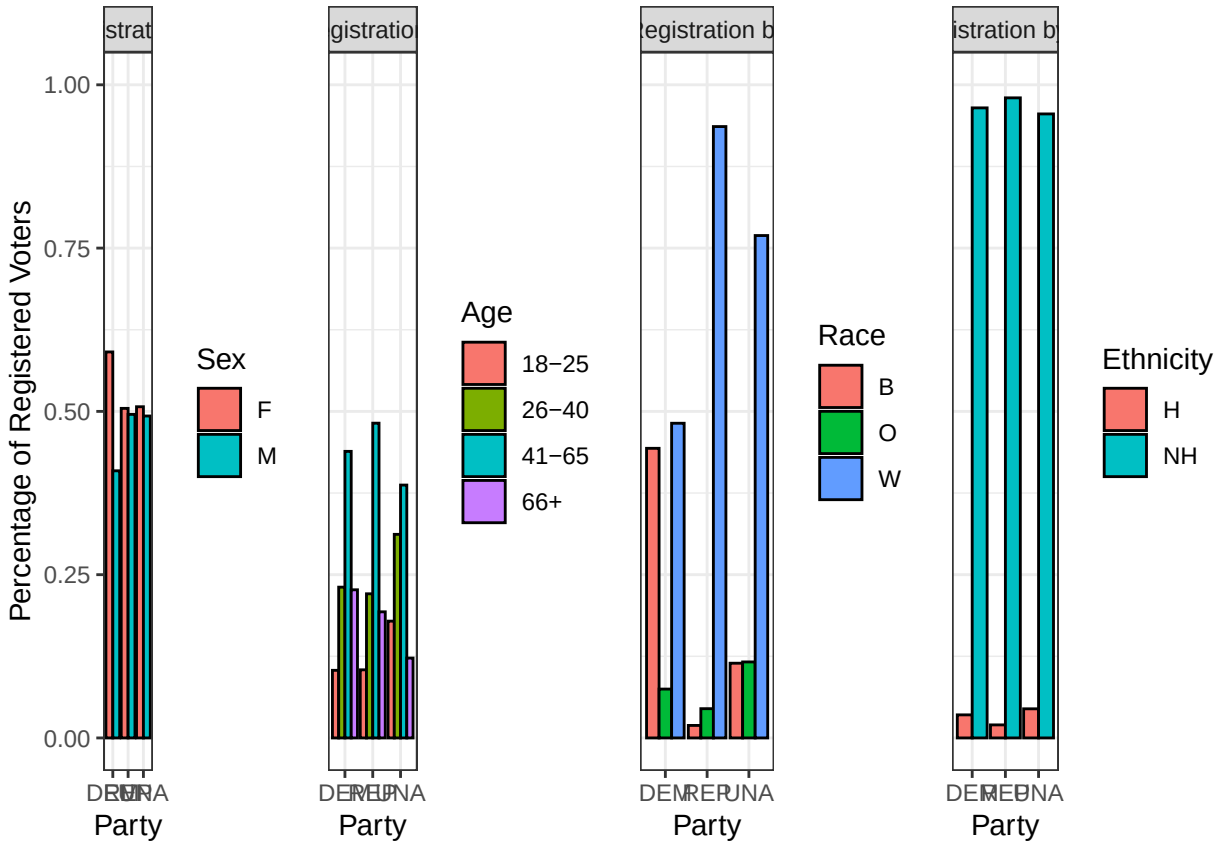

```
hisp2 <- df_select %>%
  group_by(ethnicity, party_cd) %>%
  summarize(total_voters = sum(total_voters)) %>%
  ungroup() %>%
  group_by(party_cd) %>%
  mutate(percentage = total_voters/sum(total_voters))
```

```
## `summarise()` has grouped output by 'ethnicity'. You can override using the
## `.groups` argument.
```

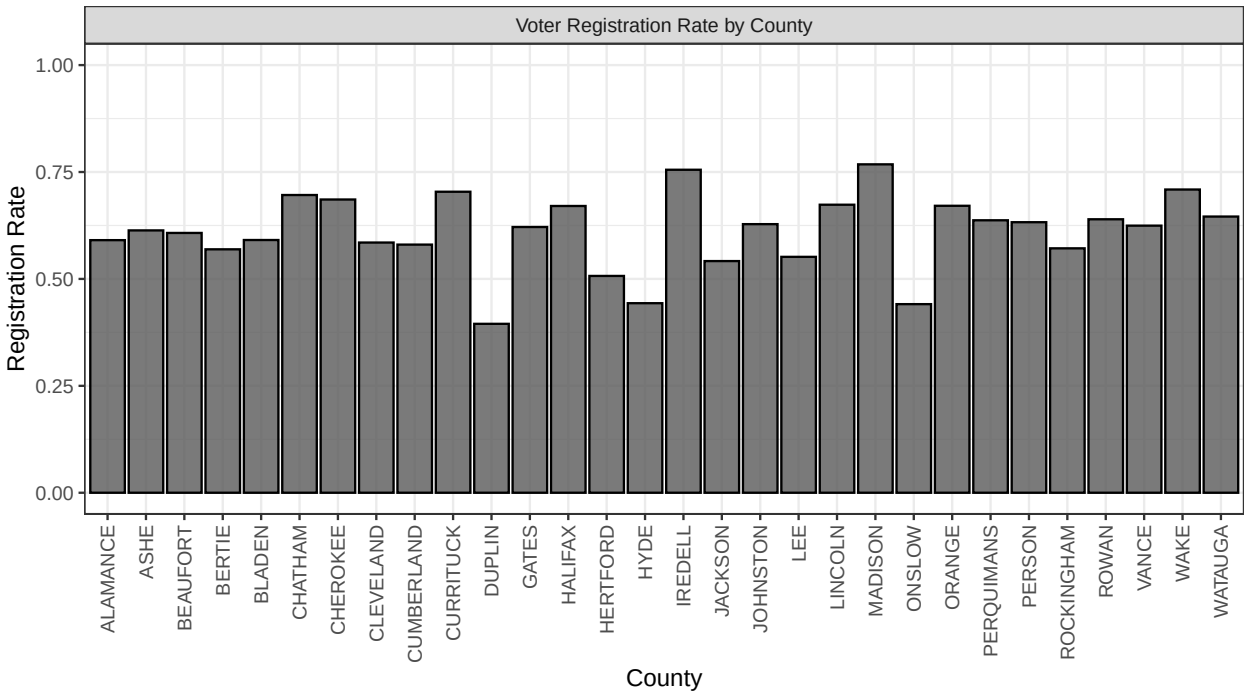
```
eth_plot <- ggplot(data = hisp2, aes(x=party_cd, y=percentage,
                                     fill = ethnicity)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  ylim(0, 1) +
  theme_bw() +
  labs(x="Party", y="Percentage of Registered Voters", fill = "Ethnicity") +
  scale_fill_discrete(labels=c("H", "NH")) +
  facet_wrap(~"Party Registration by Ethnicity")

eth_plot
```




```
remove_y <- theme(
  axis.text.y = element_blank(),
  axis.ticks.y = element_blank(),
  axis.title.y = element_blank())

ggarrange(sex_plot, age_plot + remove_y, race_plot + remove_y, eth_plot + remove_y, nrow = 1)
```




```
# Explore County

county_pct <- df_select %>%
  group_by(county_desc) %>%
  summarise(x = sum(total_voters)/sum(Freq))


ggplot(county_pct, aes(x = county_desc, y = x)) +
  geom_bar(stat = "identity", color = "black", alpha = .8) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ylim(0, 1) +
  facet_wrap(~"Voter Registration Rate by County") +
  labs(x= "County", y= "Registration Rate")
```




```
df_select %>%
  filter(county_desc %in% c("DUPLIN", "HYDE", "ONSLOW", "IREDELL", "MADISON")) %>%
  group_by(county_desc, party_cd) %>%
  summarise(sum(total_voters))
```

```
## `summarise()` has grouped output by 'county_desc'. You can override using the
## `.groups` argument.
```

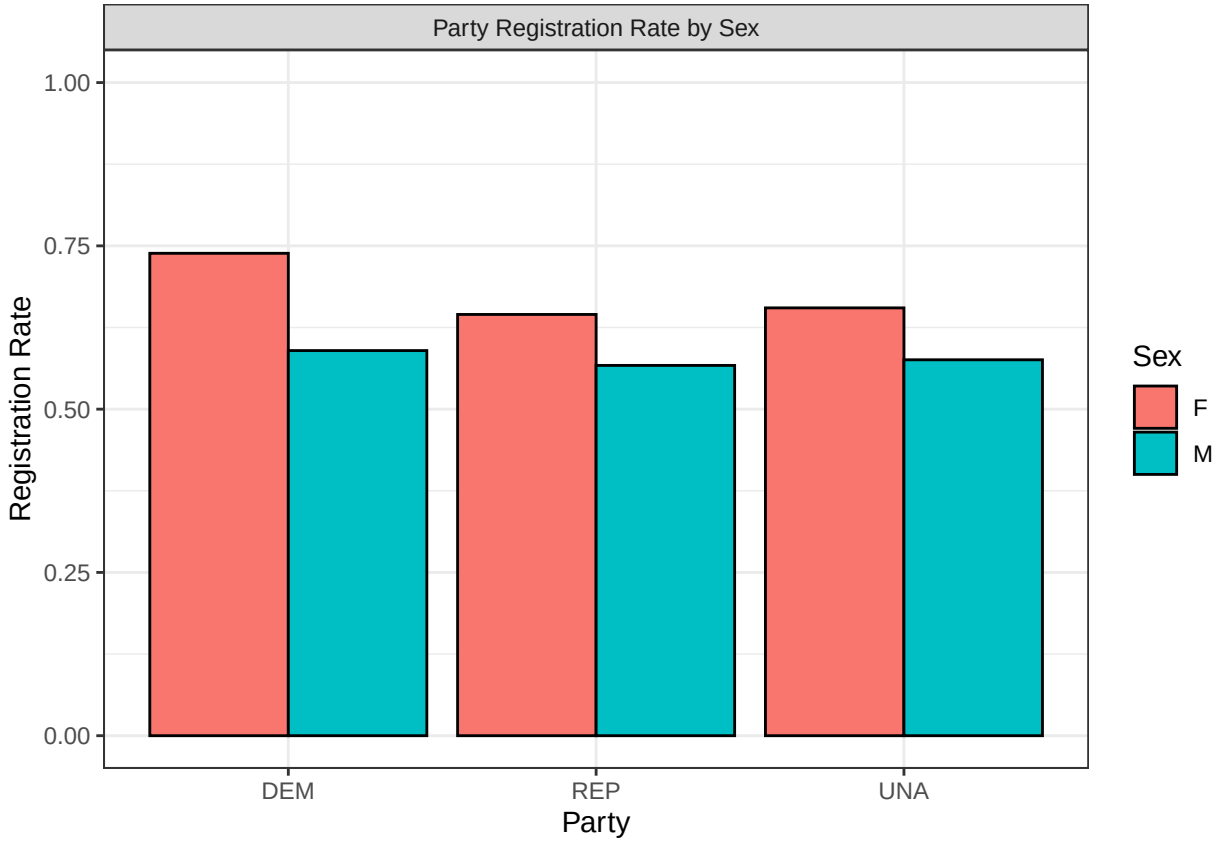
```
## # A tibble: 15 x 3
## # Groups:   county_desc [5]
##    county_desc party_cd `sum(total_voters)`
##    <chr>       <chr>                  <dbl>
##  1 DUPLIN      DEM                    14116
##  2 DUPLIN      REP                     7993
##  3 DUPLIN      UNA                     7135
##  4 HYDE        DEM                     2025
##  5 HYDE        REP                      523
##  6 HYDE        UNA                      823
##  7 IREDELL     DEM                    30863
##  8 IREDELL     REP                    48459
##  9 IREDELL     UNA                    36504
## 10 MADISON     DEM                     6700
## 11 MADISON     REP                     4318
## 12 MADISON     UNA                     5391
## 13 ONSLOW      DEM                    29114
## 14 ONSLOW      REP                    37602
## 15 ONSLOW      UNA                    37009
```

```
# Sex within affiliation
sex_aff <- df_select %>%
```

```
  group_by(sex, party_cd) %>%
  summarise(x = sum(total_voters)/sum(Freq))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

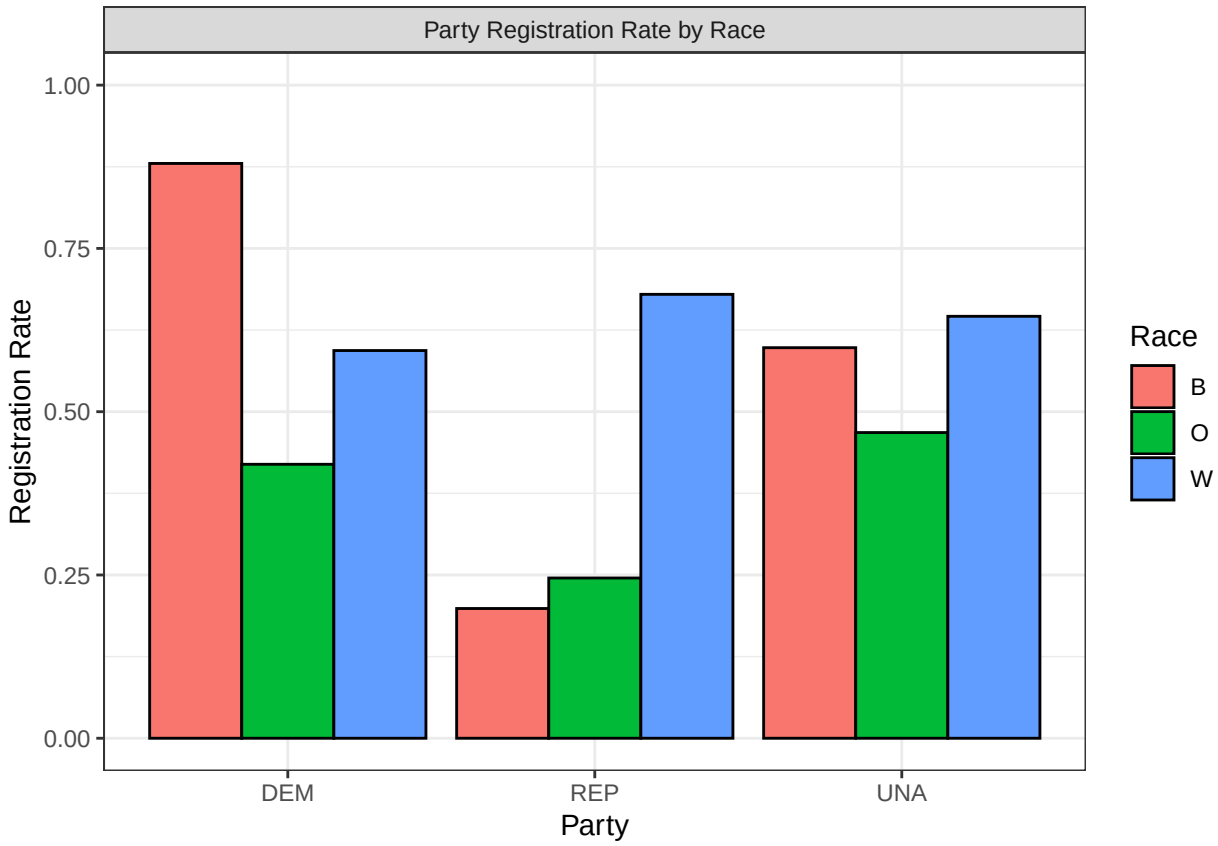
```
sexplot2 <- ggplot(data = sex_aff, aes(x=party_cd, y=x, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  scale_fill_discrete(labels=c("F", "M")) +
  facet_wrap(~"Party Registration Rate by Sex") +
  labs(x= "Party", y= "Registration Rate", fill = "Sex")

sexplot2
```




```
# Race within affiliation
race_aff <- df_select %>%
  group_by(race, party_cd) %>%
  summarise(x = sum(total_voters)/sum(Freq))
```

```
## `summarise()` has grouped output by 'race'. You can override using the
## `.groups` argument.
```

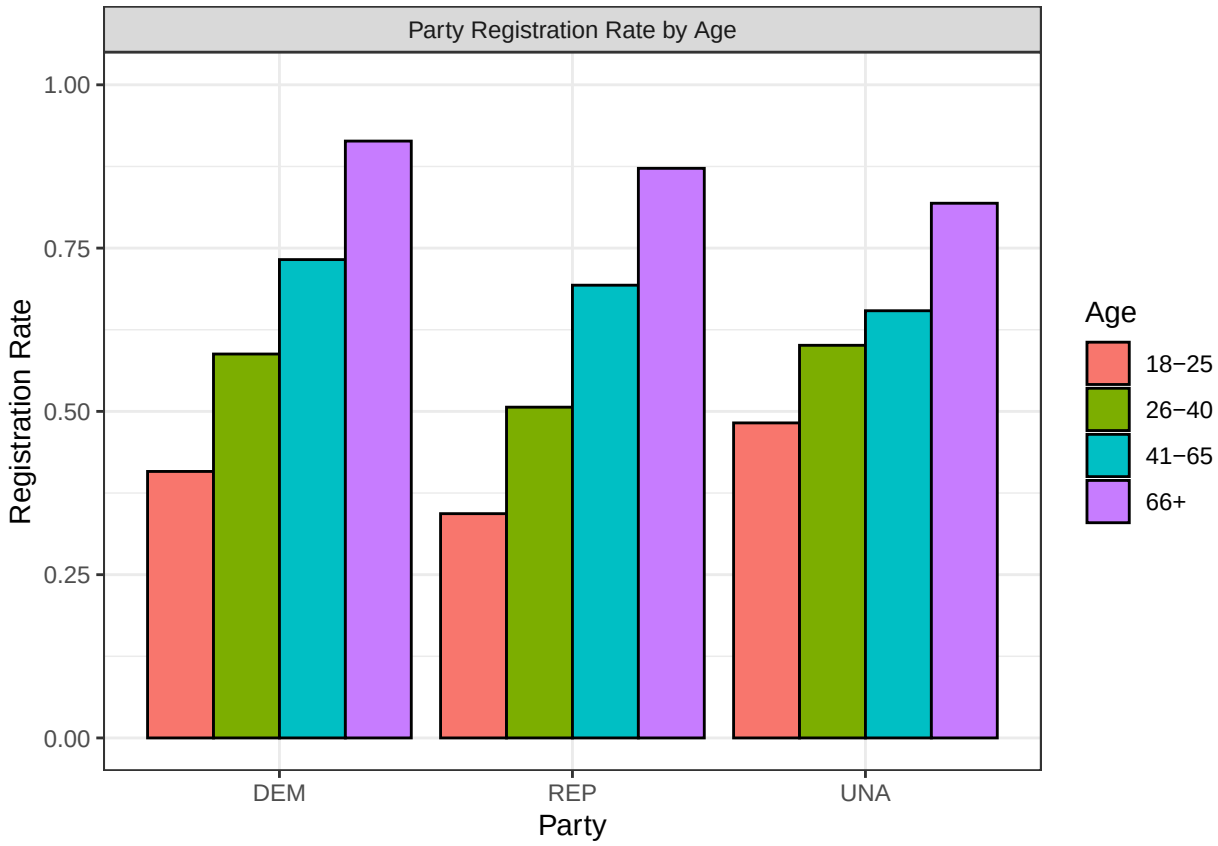
```
raceplot2 <- ggplot(data = race_aff, aes(x=party_cd, y=x, fill = race)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  scale_fill_discrete(labels=c("B", "O", "W")) +
  facet_wrap(~"Party Registration Rate by Race") +
  labs(x= "Party", y= "Registration Rate", fill = "Race")

raceplot2
```




```
# Age within affiliation
age_aff <- df_select %>%
  group_by(age, party_cd) %>%
  summarise(x = sum(total_voters)/sum(Freq))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

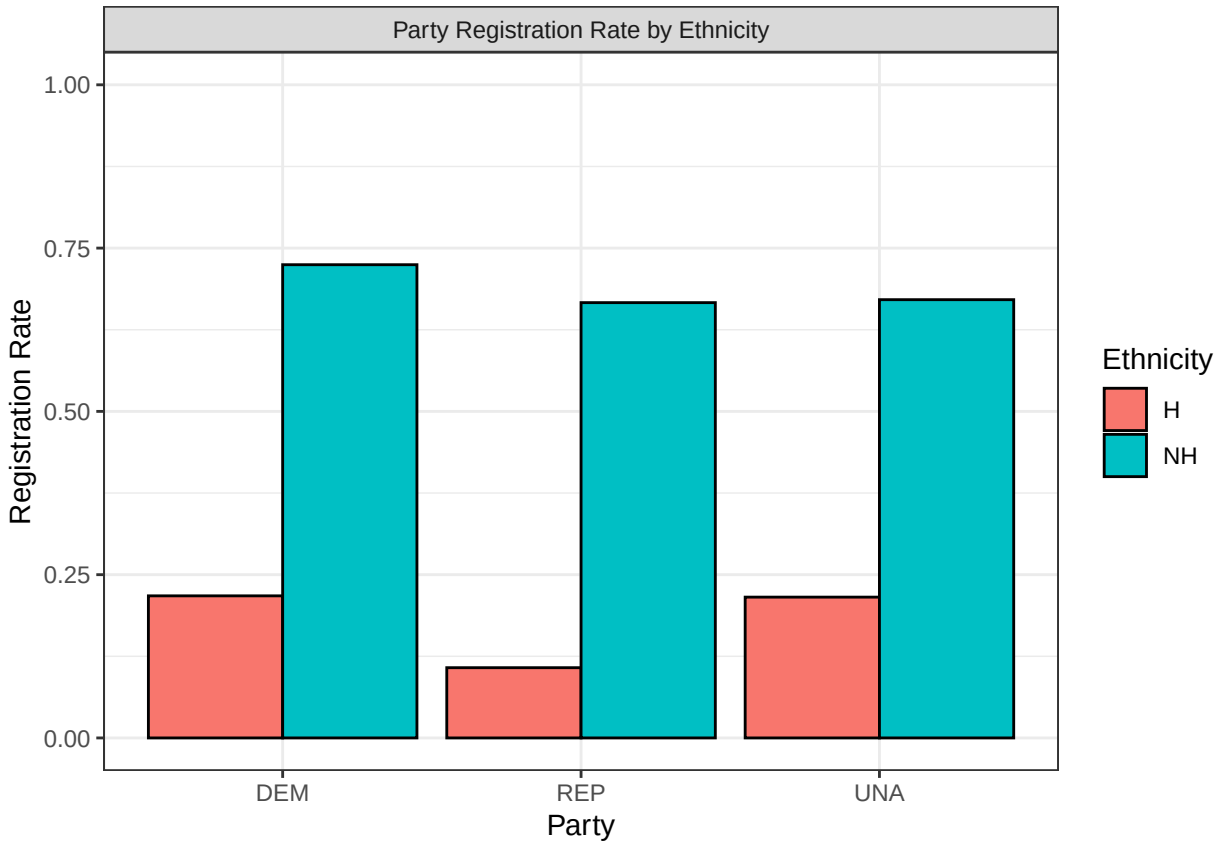
```
ageplot2 <- ggplot(data = age_aff, aes(x=party_cd, y=x, fill = age)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  facet_wrap(~"Party Registration Rate by Age") +
  labs(x= "Party", y= "Registration Rate", fill = "Age")

ageplot2
```




```
# Ethnicity within affiliation
eth_aff <- df_select %>%
  group_by(ethnicity, party_cd) %>%
  summarise(x = sum(total_voters)/sum(Freq))
```

```
## `summarise()` has grouped output by 'ethnicity'. You can override using the
## `.groups` argument.
```

```
ethplot2 <- ggplot(data = eth_aff, aes(x=party_cd, y=x, fill = ethnicity)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  theme_bw() +
  ylim(0, 1) +
  scale_fill_discrete(labels=c("H", "NH")) +
  facet_wrap(~"Party Registration Rate by Ethnicity") +
  labs(x= "Party", y= "Registration Rate", fill = "Ethnicity")

ethplot2
```




```
# Modeling

set.seed(45)


fit3 <- brm(data = df_select, family = binomial,
            formula = total_voters | trials(Freq) ~ sex + race + party_cd + age +
                      ethnicity + (1| county_desc),
            prior = c(prior(normal(0, 10), class = Intercept),
            prior(normal(0, 1), class = b),
            prior(cauchy(0, 1), class = sd)),
            iter = 2500, warmup = 500, cores = 2, chains = 2,seed = 10)
```

```
## Compiling Stan program...

## Start sampling

## Warning: There were 103 transitions after warmup that exceeded the maximum treedepth. Increase max_t
## https://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded

## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
summary(fit3)
```

```
##  Family: binomial
##   Links: mu = logit
## Formula: total_voters | trials(Freq) ~ sex + race + party_cd + age + ethnicity + (1 | county_desc)
##    Data: df_select (Number of observations: 3870)
##   Draws: 2 chains, each with iter = 2500; warmup = 500; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~county_desc (Number of levels: 30)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.48      0.07     0.37     0.63 1.00      381      604
##
## Population-Level Effects:
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                -1.58      0.09    -1.75    -1.41 1.00      378
## sexMale                  -0.40      0.00    -0.41    -0.40 1.00     4740
## raceOther                -0.86      0.01    -0.87    -0.85 1.00     2537
## raceWhite                -0.62      0.00    -0.63    -0.61 1.00     2928
## party_cdREP              -0.14      0.00    -0.14    -0.13 1.00     3763
## party_cdUNA               0.01      0.00     0.00     0.02 1.00     3804
## age26M40                  0.71      0.00     0.70     0.72 1.00     2838
## age41M65                  1.09      0.00     1.08     1.09 1.00     2879
## age66P                    2.22      0.01     2.21     2.23 1.00     2665
## ethnicityNonMHispanic     2.07      0.00     2.06     2.08 1.00     2838
##                       Tail_ESS
## Intercept                  502
## sexMale                   2820
## raceOther                 1746
## raceWhite                 3024
## party_cdREP               2631
## party_cdUNA               2865
## age26M40                  2754
## age41M65                  2340
## age66P                    2761
## ethnicityNonMHispanic     2263
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

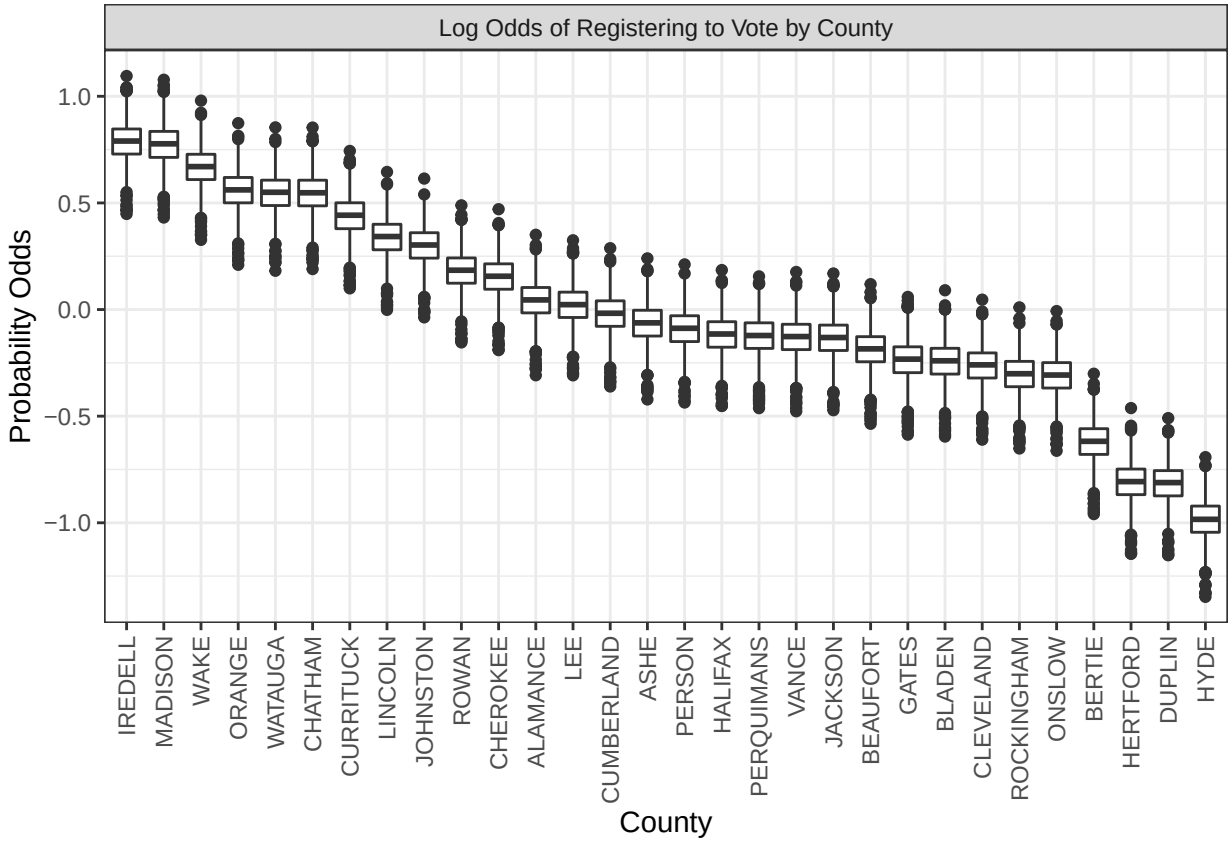
```
post <- posterior_samples(fit3)
```

```
## Warning: Method 'posterior_samples' is deprecated. Please see ?as_draws for
## recommended alternatives.
```

```
# County model output

county_hist <- post[,12:41]

melt_county_hist <- melt(county_hist)
```

```
## Using  as id variables
```

```
melt_county_hist$variable <- str_sub(melt_county_hist$variable, 15, str_length(melt_county_hist$variable

ggplot(melt_county_hist, aes(x = reorder(variable, -value), y = value)) +
  geom_boxplot() +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(x="County", y="Probability Odds") +
  facet_wrap(~"Log Odds of Registering to Vote by County")
```




```
# model fit/posterior predictive checks
set.seed(45)
out <- predict(fit3, summary = FALSE)

t(out)[5, 1:10]
```

```
##  [1] 412 420 423 423 420 416 421 422 415 433
```

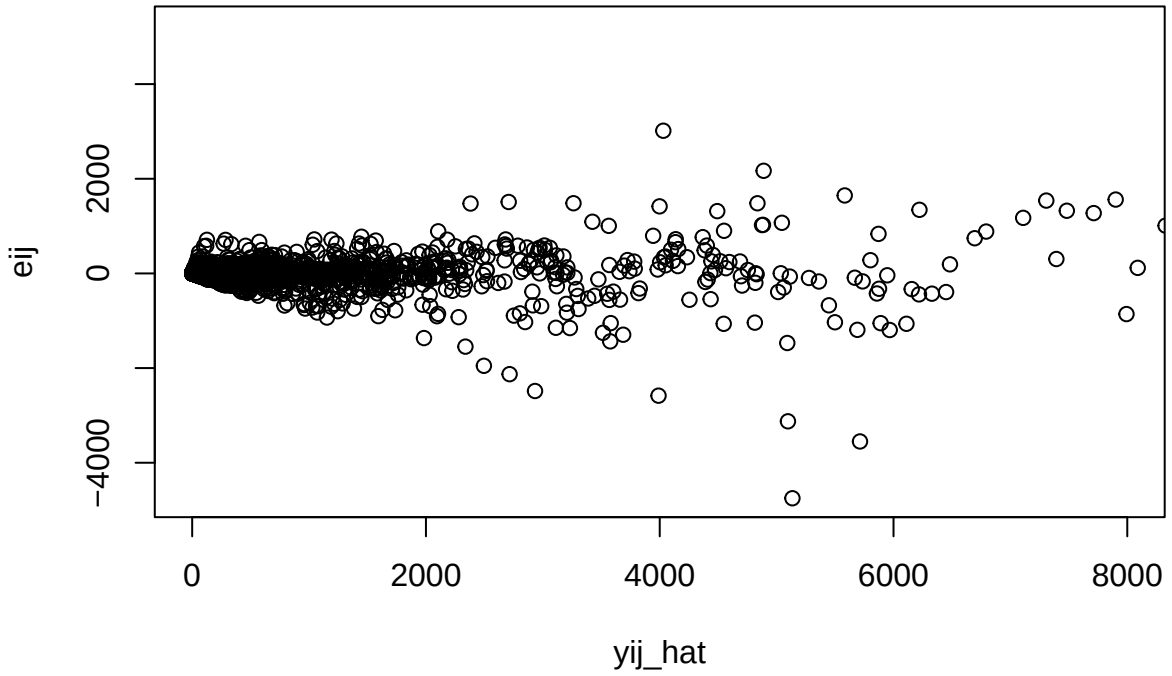
```
df_select[5,]
```

```
## # A tibble: 1 x 10
## # Groups:   county_desc, race, age, sex, ethnicity [1]
##   county_desc race  age   sex    ethnicity   party~1 total~2  Freq Total~3 Freq2
##   <chr>       <chr> <chr> <chr>  <chr>       <chr>     <dbl> <dbl>   <dbl> <dbl>
## 1 ALAMANCE    Black 18-25 Female Non-Hispan~ UNA         283   657  151131  1648
## # ... with abbreviated variable names 1: party_cd, 2: total_voters,
## #   3: TotalCountyPopulation
```

```
nrow(df_select)
```

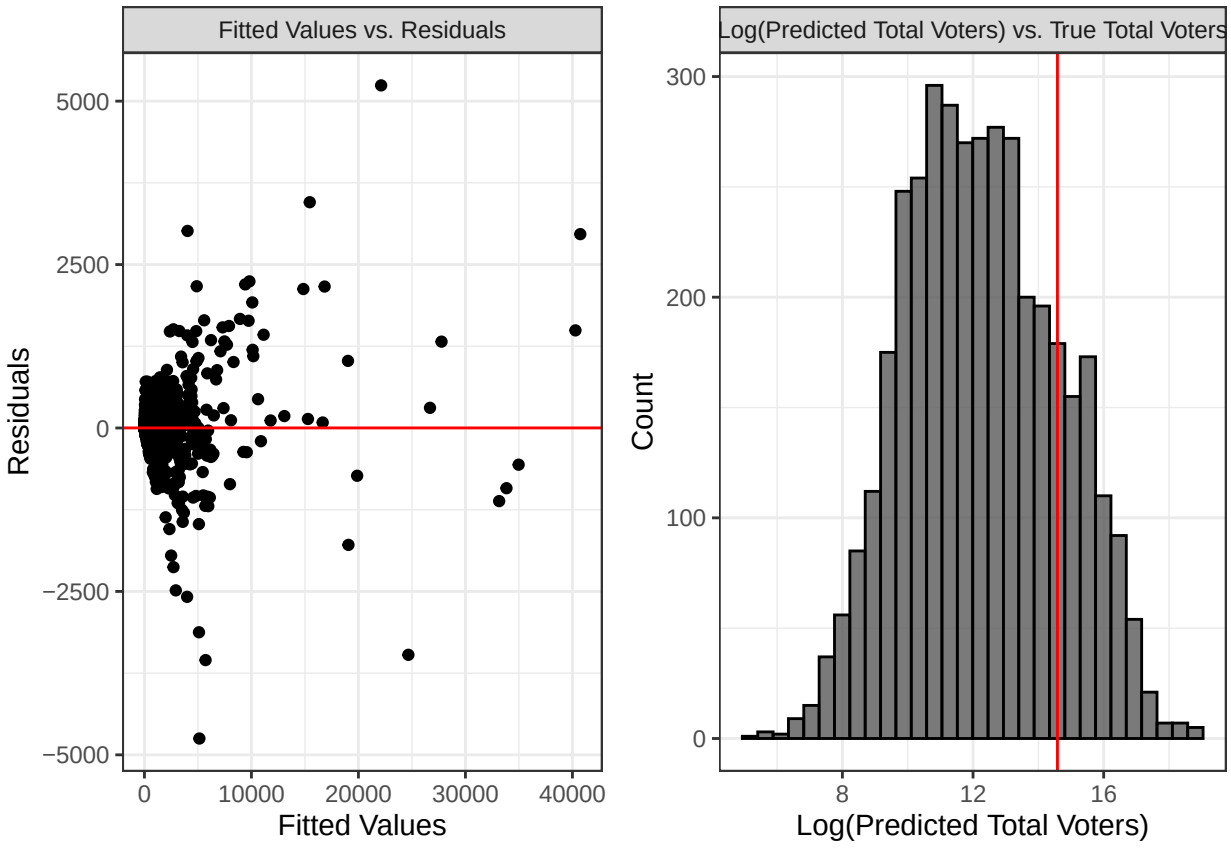
```
## [1] 3870
```

```
yij_hat <- colMeans(out)
eij <- df_select$total_voters - yij_hat

resid_m1 <- as.data.frame(cbind(yij_hat, eij))

plot(yij_hat, eij, xlim = c(0, 8000))
```




```
res <- ggplot(data = resid_m1, aes(x = yij_hat, y = eij)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 1, color = "red") +
  labs(x="Fitted Values", y="Residuals") +
  facet_wrap(~"Fitted Values vs. Residuals")


sim_all <- as.data.frame(colSums(out))
tot_all <- sum(df_select$total_voters)


m1fit <- ggplot(data = sim_all, aes(x=log(`colSums(out)`))) +
  geom_histogram(color = "black", alpha = .8) +
  geom_vline(xintercept = log(tot_all), color = "red") +
  labs(x="Log(Predicted Total Voters)", y="Count") +
  facet_wrap(~"Log(Predicted Total Voters) vs. True Total Voters") +
  theme_bw()

ggarrange(res, m1fit, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# republican model
set.seed(45)
fit_r <-  brm(data = df_select %>% filter(party_cd == "REP"), family = binomial,
            formula = total_voters | trials(Freq) ~ 1 + sex + race + age + ethnicity +
              (1 | county_desc),
            prior = c(prior(normal(0, 10), class = Intercept),
            prior(normal(0, 1), class = b),
            prior(cauchy(0,1), class = sd)),
            iter = 2500, warmup = 500, cores = 2, chains = 2,seed = 10)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

```
summary(fit_r)
```

```
##  Family: binomial
##   Links: mu = logit
## Formula: total_voters | trials(Freq) ~ 1 + sex + race + age + ethnicity + (1 | county_desc)
##    Data: df_select %>% filter(party_cd == "REP") (Number of observations: 1175)
##   Draws: 2 chains, each with iter = 2500; warmup = 500; thin = 1;
##          total post-warmup draws = 4000
##
```

```
## Group-Level Effects:
## ~county_desc (Number of levels: 30)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.55      0.07     0.42     0.71 1.00      611      772
##
## Population-Level Effects:
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                -4.79      0.10    -4.99    -4.61 1.00      415
## sexMale                  -0.26      0.00    -0.27    -0.25 1.00     4142
## raceOther                 1.00      0.01     0.97     1.02 1.00     2471
## raceWhite                 2.02      0.01     1.99     2.04 1.00     2504
## age26M40                  0.79      0.01     0.78     0.80 1.00     2596
## age41M65                  1.34      0.01     1.33     1.35 1.00     2807
## age66P                    2.31      0.01     2.29     2.33 1.00     2805
## ethnicityNonMHispanic     2.50      0.01     2.48     2.52 1.00     3522
##                       Tail_ESS
## Intercept                  669
## sexMale                   2450
## raceOther                 1871
## raceWhite                 2582
## age26M40                  2703
## age41M65                  2728
## age66P                    2673
## ethnicityNonMHispanic     2865
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
# democrat model
set.seed(45)

fit_d <-  brm(data = df_select %>% filter(party_cd == "DEM"), family = binomial,
            formula = total_voters | trials(Freq) ~ 1 + sex + race + age + ethnicity +
              (1 | county_desc),
            prior = c(prior(normal(0, 10), class = Intercept),
            prior(normal(0, 1), class = b),
            prior(cauchy(0,1), class = sd)),
            iter = 2500, warmup = 500, cores = 2, chains = 2,seed = 10)
```

```
## Compiling Stan program...

## Start sampling
```

```
summary(fit_d)
```

```
##  Family: binomial
##   Links: mu = logit
## Formula: total_voters | trials(Freq) ~ 1 + sex + race + age + ethnicity + (1 | county_desc)
##    Data: df_select %>% filter(party_cd == "DEM") (Number of observations: 1377)
##   Draws: 2 chains, each with iter = 2500; warmup = 500; thin = 1;
##          total post-warmup draws = 4000
```

```
##
## Group-Level Effects:
## ~county_desc (Number of levels: 30)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.47      0.06     0.36     0.61 1.00      503      619
##
## Population-Level Effects:
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                -0.55      0.09    -0.72    -0.38 1.00      464
## sexMale                  -0.61      0.00    -0.62    -0.60 1.00     4085
## raceOther                -1.65      0.01    -1.67    -1.64 1.00     2644
## raceWhite                -1.86      0.01    -1.87    -1.85 1.00     2969
## age26M40                  0.81      0.01     0.79     0.82 1.00     2449
## age41M65                  1.35      0.01     1.34     1.36 1.00     2437
## age66P                    2.86      0.01     2.84     2.88 1.00     2509
## ethnicityNonMHispanic     1.83      0.01     1.81     1.84 1.00     2939
##                       Tail_ESS
## Intercept                  803
## sexMale                   2791
## raceOther                 2676
## raceWhite                 2833
## age26M40                  2339
## age41M65                  2420
## age66P                    2529
## ethnicityNonMHispanic     2575
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
#unaffiliated model
set.seed(45)

fit_u <-  brm(data = df_select %>% filter(party_cd == "UNA"), family = binomial,
            formula = total_voters | trials(Freq) ~ 1 + sex + race + age + ethnicity +
              (1 | county_desc),
            prior = c(prior(normal(0, 10), class = Intercept),
            prior(normal(0, 1), class = b),
            prior(cauchy(0,1), class = sd)),
            iter = 2500, warmup = 500, cores = 2, chains = 2,seed = 10)
```

```
## Compiling Stan program...

## Start sampling

## Warning: There were 4 transitions after warmup that exceeded the maximum treedepth. Increase max_tre
## https://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded

## Warning: Examine the pairs() plot to diagnose sampling problems
```

```
summary(fit_u)
```

```
##  Family: binomial
##   Links: mu = logit
## Formula: total_voters | trials(Freq) ~ 1 + sex + race + age + ethnicity + (1 | county_desc)
##    Data: df_select %>% filter(party_cd == "UNA") (Number of observations: 1318)
##   Draws: 2 chains, each with iter = 2500; warmup = 500; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~county_desc (Number of levels: 30)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.56      0.08     0.43     0.73 1.00      530      882
##
## Population-Level Effects:
##                       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                -1.97      0.10    -2.17    -1.77 1.00      472
## sexMale                  -0.32      0.00    -0.33    -0.31 1.00     4196
## raceOther                 0.00      0.01    -0.01     0.02 1.00     2660
## raceWhite                 0.05      0.01     0.04     0.07 1.00     2948
## age26M40                  0.55      0.01     0.54     0.56 1.00     2877
## age41M65                  0.63      0.01     0.62     0.64 1.00     2828
## age66P                    1.45      0.01     1.43     1.47 1.00     2877
## ethnicityNonMHispanic     2.04      0.01     2.02     2.05 1.00     3391
##                       Tail_ESS
## Intercept                  582
## sexMale                   2512
## raceOther                 2467
## raceWhite                 2494
## age26M40                  2915
## age41M65                  2358
## age66P                    2691
## ethnicityNonMHispanic     2733
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

```
# sex and race posterior densisties
set.seed(45)
samp_r <- posterior_samples(fit_r)
```

```
## Warning: Method 'posterior_samples' is deprecated. Please see ?as_draws for
## recommended alternatives.
```

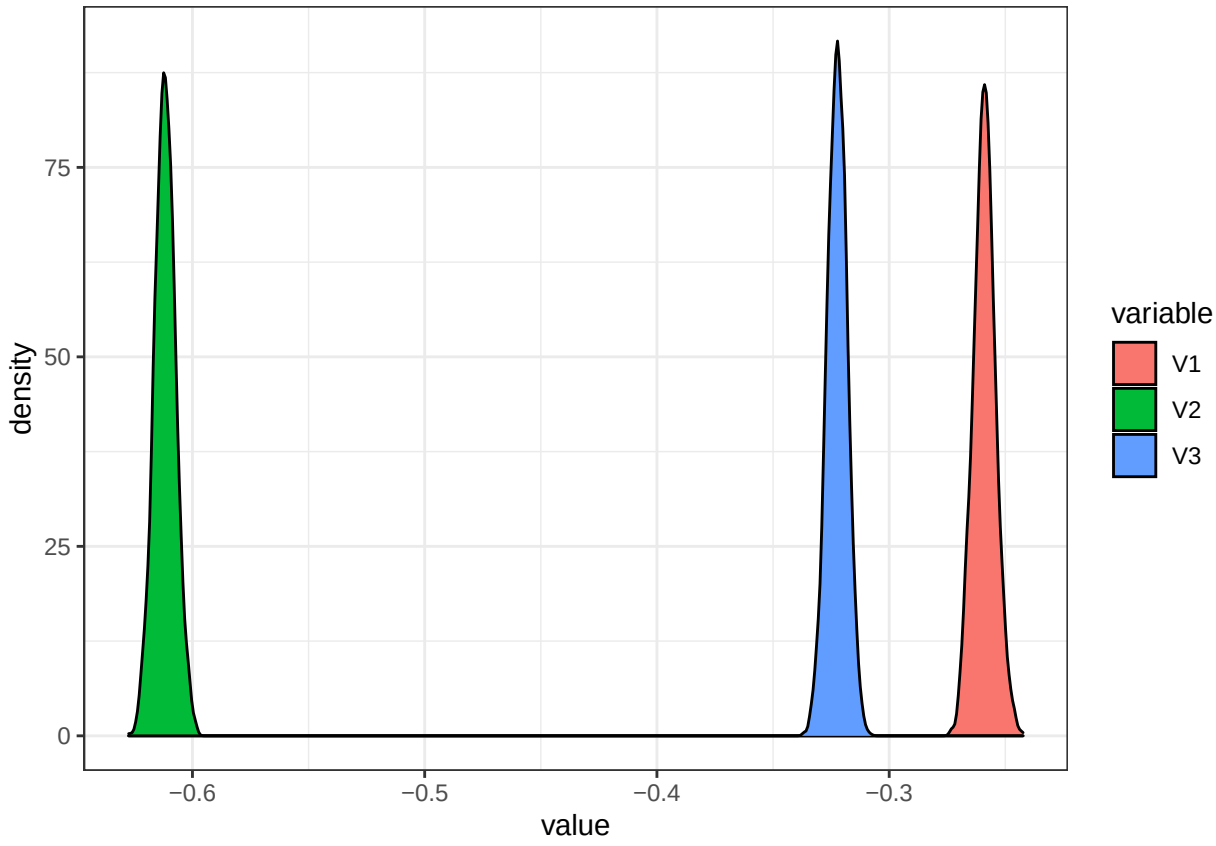
```
samp_d <- posterior_samples(fit_d)
```

```
## Warning: Method 'posterior_samples' is deprecated. Please see ?as_draws for
## recommended alternatives.
```

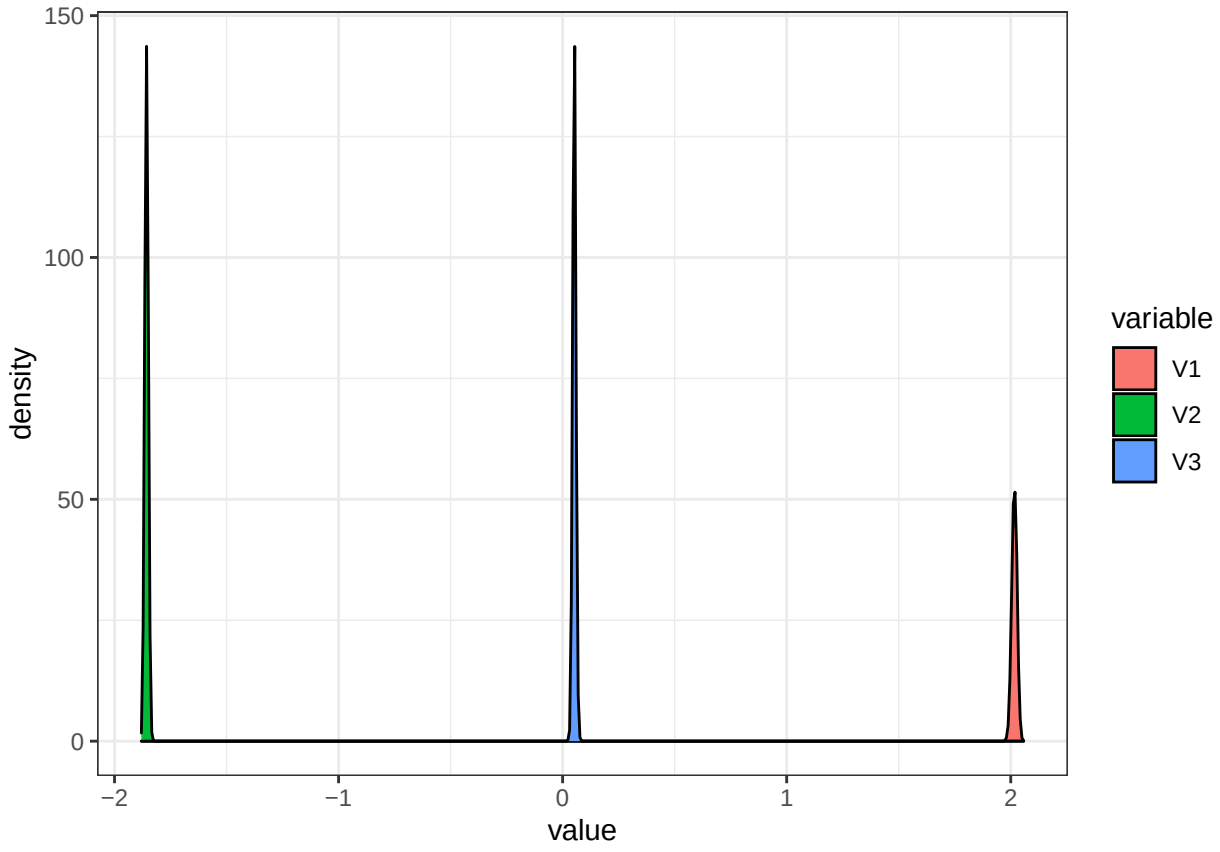
```
samp_u <- posterior_samples(fit_u)
```

```
## Warning: Method 'posterior_samples' is deprecated. Please see ?as_draws for
## recommended alternatives.
```

```
samp_all_s <- as.data.frame(cbind(samp_r$b_sexMale, samp_d$b_sexMale, samp_u$b_sexMale))

melt_samp_s <- melt(samp_all_s)

## Using  as id variables

ggplot(melt_samp_s, aes(x=value, fill=variable)) +
  geom_density() +
  theme_bw()
```



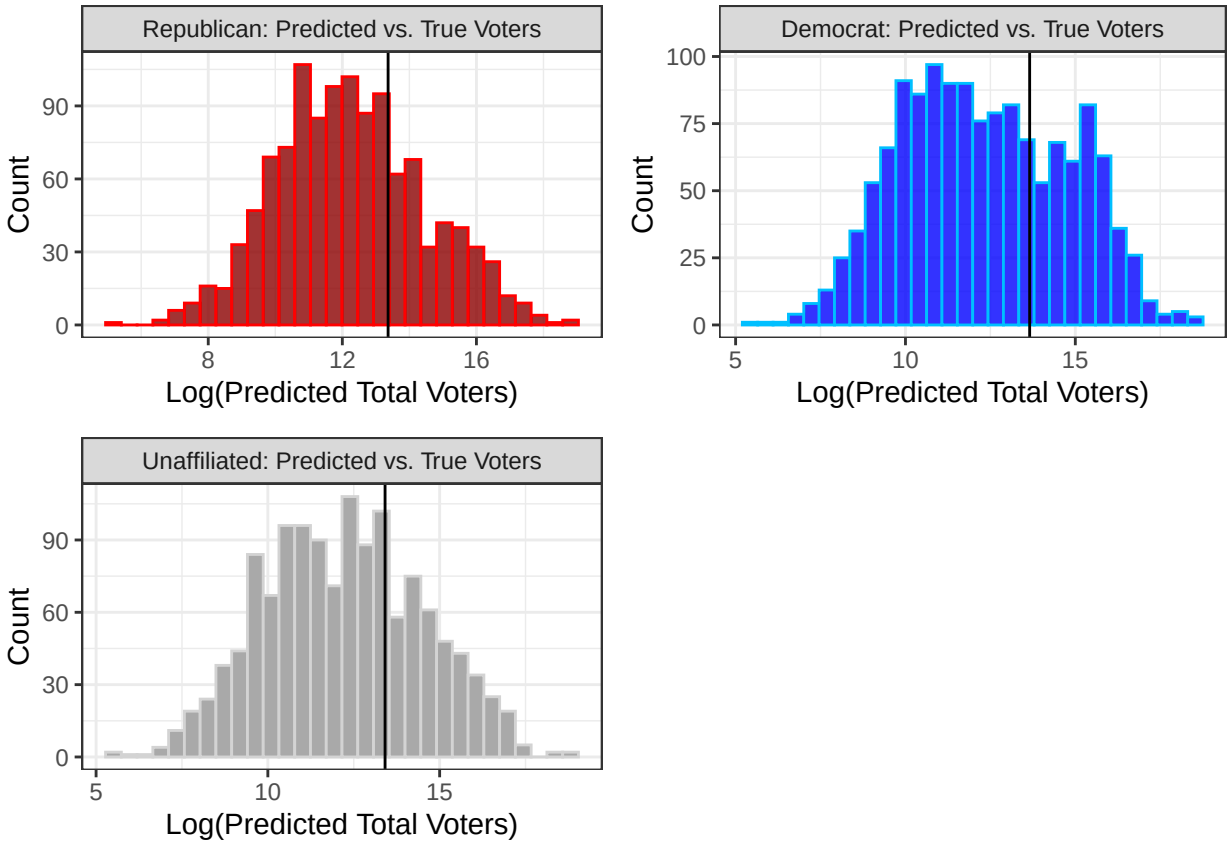

```
samp_all_r <- as.data.frame(cbind(samp_r$b_raceWhite, samp_d$b_raceWhite, samp_u$b_raceWhite))

melt_samp_r <- melt(samp_all_r)

## Using  as id variables

ggplot(melt_samp_r, aes(x=value, fill=variable)) +
  geom_density() +
  theme_bw()
```




```
# model fit/posterior predictive checks

id_rep <- which(df_select$party_cd == "REP")
id_dem <- which(df_select$party_cd == "DEM")
id_una <- which(df_select$party_cd == "UNA")


sim_rep <- as.data.frame(colSums(out[,id_rep]))
tot_rep <- sum(df_select$total_voters[id_rep])

r <- ggplot(data = sim_rep, aes(x=log(`colSums(out[, id_rep])`))) +
  geom_histogram(color = "red1", fill = "red4", alpha = .8) +
  geom_vline(xintercept = log(tot_rep), color = "black") +
  theme_bw() +
  labs(x="Log(Predicted Total Voters)", y="Count") +
  facet_wrap(~"Republican: Predicted vs. True Voters")

sim_dem <- as.data.frame(colSums(out[,id_dem]))
tot_dem <- sum(df_select$total_voters[id_dem])

d <- ggplot(data = sim_dem, aes(x=log(`colSums(out[, id_dem])`))) +
  geom_histogram(color = "deepskyblue", fill = "blue", alpha = .8) +
  geom_vline(xintercept = log(tot_dem), color = "black") +
  theme_bw() +
  labs(x="Log(Predicted Total Voters)", y="Count") +
  facet_wrap(~"Democrat: Predicted vs. True Voters")
```

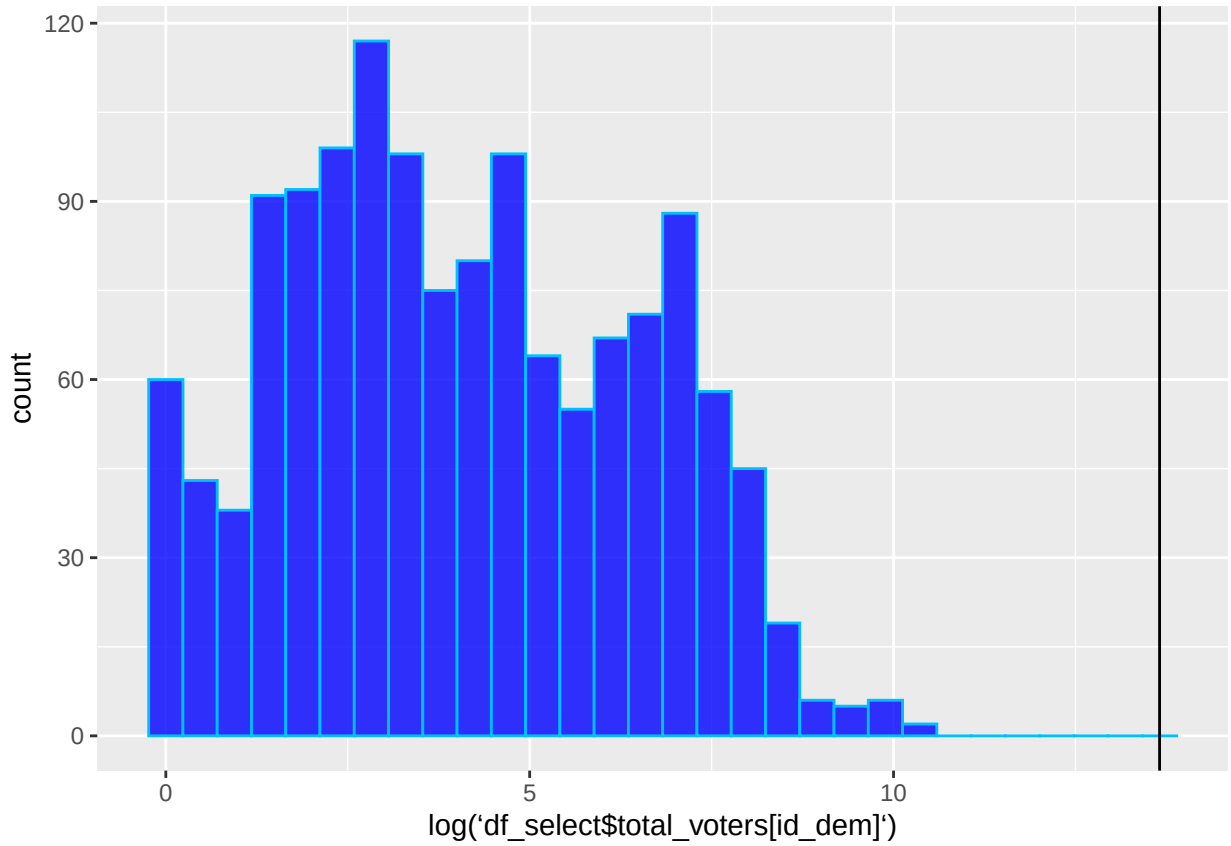
```
sim_una <- as.data.frame(colSums(out[,id_una]))
tot_una <- sum(df_select$total_voters[id_una])

u <- ggplot(data = sim_una, aes(x=log(`colSums(out[, id_una])`))) +
  geom_histogram(color = "lightgray", fill = "darkgray") +
  geom_vline(xintercept = log(tot_una), color = "black") +
  theme_bw() +
  labs(x="Log(Predicted Total Voters)", y="Count") +
  facet_wrap(~"Unaffiliated: Predicted vs. True Voters")

ggarrange(r, d, u)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = as.data.frame(df_select$total_voters[id_dem]), aes(x=log(`df_select$total_voters[id_dem]`
  geom_histogram(color = "deepskyblue", fill = "blue", alpha = .8) +
  geom_vline(xintercept = log(tot_dem), color = "black")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# model fit/posterior predictive checks
yij_hatr <- colMeans(out[,id_rep])
eijr <- df_select$total_voters[id_rep] - yij_hatr

resid_mr <- as.data.frame(cbind(yij_hatr, eijr))

res_r <- ggplot(data = resid_mr, aes(x = yij_hatr, y = eijr)) +
  geom_point(color = "red4", alpha = .8) +
  theme_bw() +
  geom_hline(yintercept = 1, color = "black") +
  labs(x="Fitted Values", y="Residuals") +
  facet_wrap(~"Republican: Fitted Values vs. Residuals")

yij_hatd <- colMeans(out[,id_dem])
eijd <- df_select$total_voters[id_dem] - yij_hatd

resid_md <- as.data.frame(cbind(yij_hatd, eijd))

res_d <- ggplot(data = resid_md, aes(x = yij_hatd, y = eijd)) +
  geom_point(color = "blue", alpha = .8) +
  theme_bw() +
  geom_hline(yintercept = 1, color = "black") +
  labs(x="Fitted Values", y="Residuals") +
  facet_wrap(~"Democrat: Fitted Values vs. Residuals")

yij_hatu <- colMeans(out[,id_una])
```

```r
eiju <- df_select$total_voters[id_una] - yij_hatu

resid_mu <- as.data.frame(cbind(yij_hatu, eiju))

res_u <- ggplot(data = resid_mu, aes(x = yij_hatu, y = eiju)) +
  geom_point(color = "gray") +
  theme_bw() +
  geom_hline(yintercept = 1, color = "black") +
  labs(x="Fitted Values", y="Residuals") +
  facet_wrap(~"Unaffiliated: Fitted Values vs. Residuals")

res_u
```