CS7DS3 Main Assignment

26-04-2025

Contents

Summary

This analysis investigates whether voting for Trump in the 2024 election was associated with increased rates of depression at the county level. Using data from 3,107 US counties, I found strong evidence that higher depression rates are indeed associated with greater Republican voting percentages, even after controlling for racial demographics and state-level differences. Through progressively more sophisticated modeling—from basic correlation to multiple regression to hierarchical modeling—I discovered that for each 1% increase in depression prevalence, Republican voting increases by approximately 2 percentage points. This relationship is stronger than initially calculated in simpler models. The multilevel approach revealed eye-opening state-level variation in voting patterns beyond what depression and race could really explain, with northeastern US states like Vermont and Rhode Island showing the strongest Democratic lean, while Alaska and South Dakota showed the strongest Republican lean after accounting for other factors. This suggests that though mental health factors adds a significant layer to the voting behavior, geographic and regional political cultures remain powerful influences as well.

Objective

The main task of this assignment is to analyze the USvotes data set to determine if there's evidence that voting for Trump in the 2024 election was associated with increased rates of depression. Here are the key objectives outlined to approach the task:

1. Build a statistical model to address this question
2. Identify which features associate with increased Republican votes
3. Determine if depression is associated with Republican voting after controlling for other factors
4. Pay special attention to state-level effects

Exploratory data analysis

Looking at the summary of US voters data set from Appendix B, this data set contains information about 3,107 counties across the United States. For each county, we have:

- The state and county names
- Population information (total, male, and female residents)
- Voting data from the 2024 election (total votes cast and percentage for the Republican party)
- Health information (estimated percentage of people with depression)
- Demographic information (percentage of residents who identify as white only)

Additionally, some interesting observations are made from Figure 1 and Figure 2:

1. County sizes vary dramatically - the smallest has just 43 people while the largest has over 9.6 million.
2. On average, about 67% of votes went to the Republican party across all counties, with the middle county (median) having about 70% Republican votes.
3. The Republican vote share ranges from as low as 5% in some counties to nearly 96% in others.
4. Depression rates range from about 11% to 31% across counties, with an average of about 21%.
5. The racial composition varies widely too - some counties have as low as 7% white-only residents, while others have up to 98%.
6. There's no missing data in any column, which is great for data analysis.

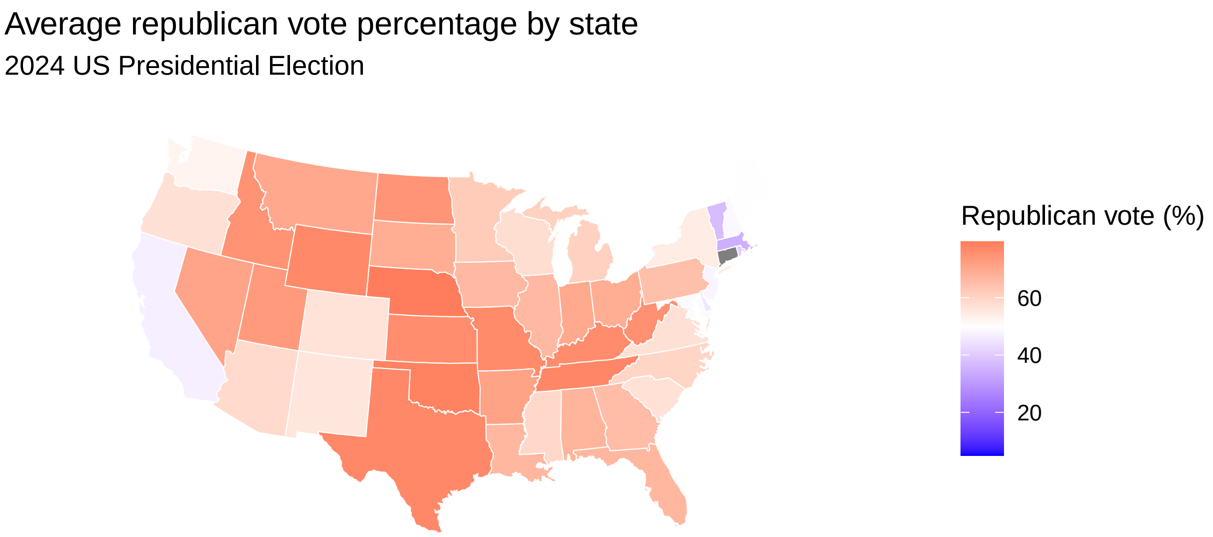
Average republican vote percentage by state
2024 US Presidential Election
Republican vote (%)
60
40
20

Figure 1: Average Republican vote percentage by state in the 2024 election; darker red indicates higher Republican share.

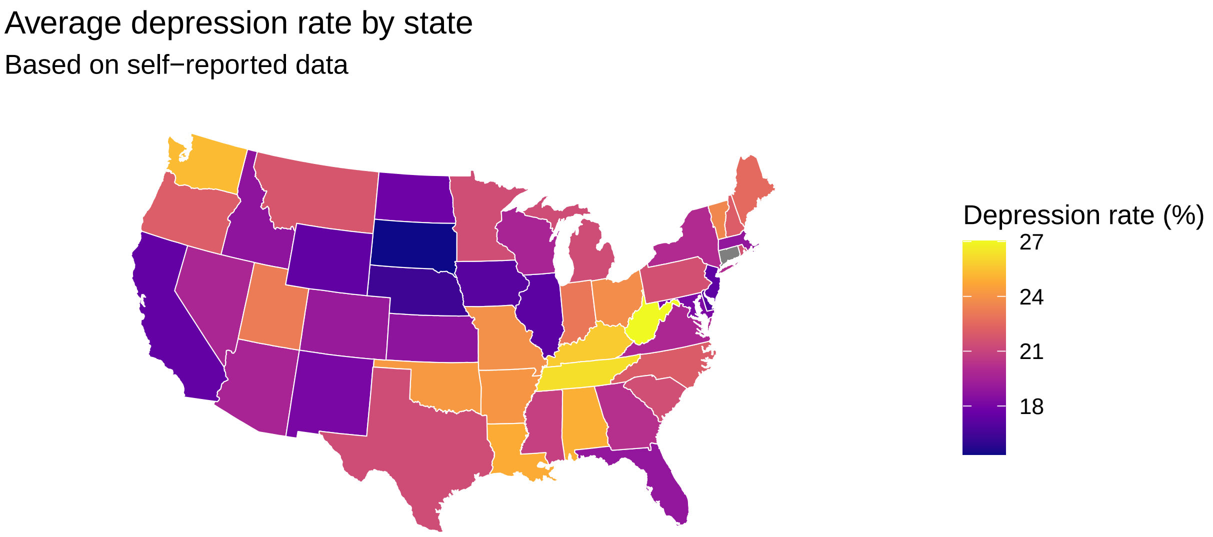

Figure 2: Average depression rate by state based on self-reported data; yellow/orange indicates higher depression rate.

Perform correlation analysis

Before building complex statistical models, I wished to run a correlation test to check for simple correlations between the multiple variables present in the US votes data set.

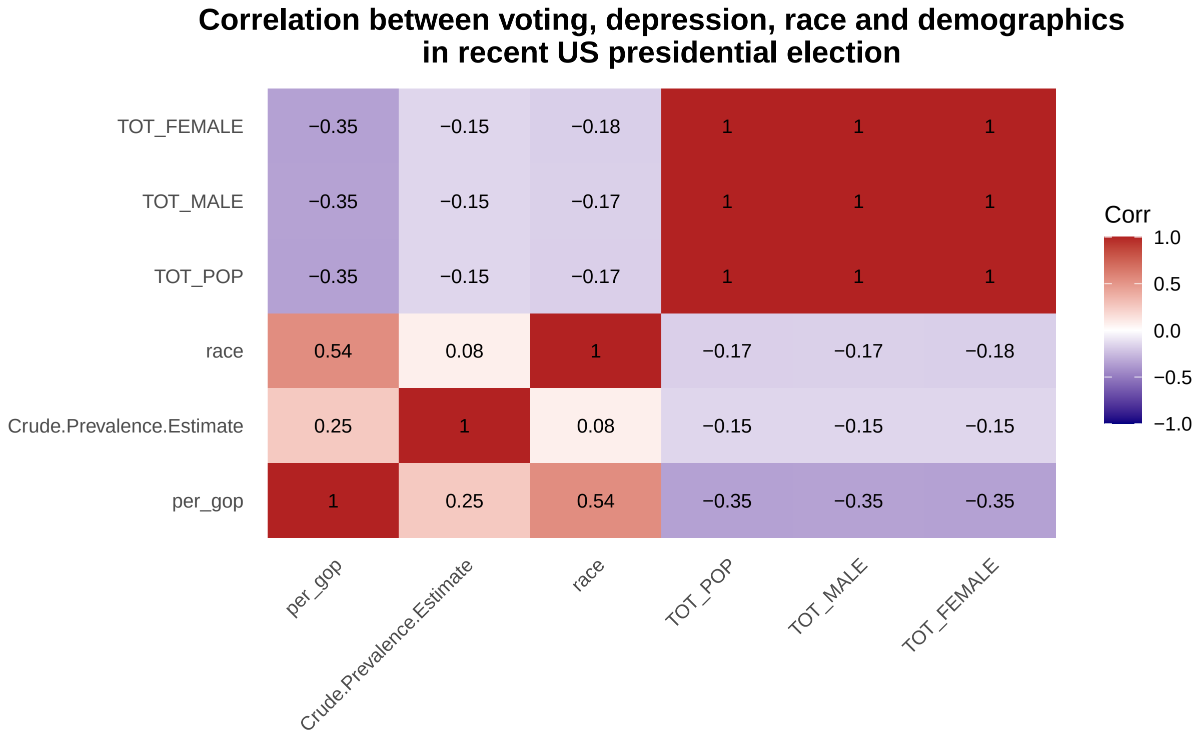

Figure 3: Plot of correlation matrix between different variables in the US votes data set.

Based on the calculated correlation matrix in Figure 3 and the specific correlation test (see Appendix E), here's what we can generally infer about the relationships in the data:

- There's a positive correlation (red box) between per_gop (Republican voting percentage) and Crude.Prevalence.Estimate (depression rates) - this is what we're primarily interested in.
- There's also a positive correlation between per_gop and race (percentage of white-only residents).
- The population variables (TOT_POP, TOT_MALE, TOT_FEMALE) are very strongly correlated with each other (blue box), which is expected.
- The population variables have slight negative correlations with per_gop (light blue boxes).

With respect to the the specific correlation test between depression and Republican voting:

- The correlation coefficient is 0.25, indicating a positive but moderate relationship
- The t-value is 14.674 with 3105 degrees of freedom
- The p-value is extremely small (2.2e-16), which is highly significant
- The 95% confidence interval for the correlation is between 0.221 and 0.287

This means that the counties with higher depression rates tend to have higher Republican voting percentages. We can say that this relationship is statistically significant (not exactly due to random chance). However, the correlation is moderate (0.25), meaning there are likely other factors influencing voting patterns. The initial analysis suggests there is evidence of a relationship between depression and Republican voting, but it's important to control for other variables (like race) and account for state-level differences before drawing strong conclusions, since correlation doesn't prove causation.

Perform linear regression

With the simple regression model (see model1 in Appendix G), we can infer from the output that depression rates are positively associated with Republican voting percentage; for each 1% increase in depression prevalence, Republican voting increases by approximately 1.28%. This relationship is statistically significant ($p < 2.2e-16$). However, depression rates alone explain only about 6.5% of the variance in Republican voting (R-squared = 0.065).

Meanwhile, in the multiple regression model (see model2 in Appendix G) with controlling for race (% white population), depression rates remain a significant predictor. For each 1% increase in depression prevalence, Republican voting increases by approximately 1.08% (slightly lower than in model1). Although, race does have a strong positive association with Republican voting (coefficient = 0.52). Overall, race and depression explain about 33.8% of the variance in Republican voting (R-squared = 0.338). This is a substantial improvement over model1, suggesting race is an important factor.

Hierarchical multilevel modeling

Based on the hierarchical model results from Figure 4 and Figure 5, there is strong evidence that depression rates are positively associated with Republican voting percentages, even after controlling for race and state-level differences. For each 1% increase in depression prevalence, Republican voting increases by approximately 2 percentage points (95% CI: 1.74 to 2.27), a stronger relationship than observed in the simpler models. The race variable remains significant with each percentage point increase in white population associated with a 0.66 percentage point increase in Republican voting. Notably, there is substantial variation between states, with a state-level standard deviation of 13.53 points, indicating important geographical differences in voting patterns beyond what depression and race explain.

The multilevel model provides a more nuanced understanding than the previous linear regressions by accounting for clustered data structure. It significantly improves model fit, reducing the residual error (sigma = 9.24) compared to the simple models. The results suggest that while both depression and racial demographics are important predictors of voting patterns, state-specific factors also play a crucial role in determining political preferences. This reinforces the idea that voting behavior is influenced by complex interactions between individual health outcomes, demographic characteristics, and regional political cultures.

Reason behind choosing hierarchical modeling

I chose to use hierarchical (multilevel) modeling for this analysis because it perfectly suits our data structure and research question. Let me explain why in simple terms.

First, our data naturally exists in layers - counties are nested within states. Regular regression treats all counties as independent, but that's not really true. Counties within the same state tend to share cultural, historical, and policy environments that influence voting patterns. Hierarchical modeling respects this natural grouping.

Second, we're specifically interested in understanding if depression rates relate to Republican voting even after accounting for state differences. Using a state-level random effect lets counties "borrow strength" from each other within states, giving us more stable estimates, especially for states with fewer counties.

Third, the multilevel approach gives us the best of both worlds - we can estimate the overall relationship between depression and voting (fixed effects) while simultaneously learning which states deviate from this

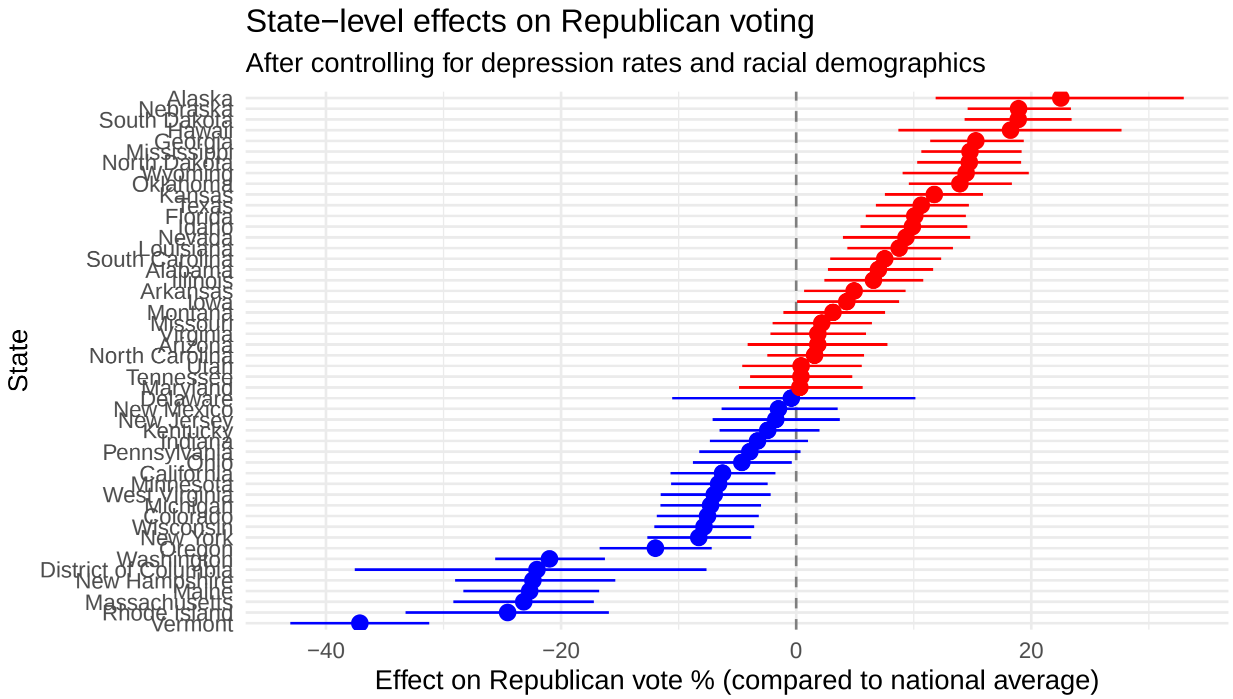

Figure 4: Estimated state-level deviations in Republican vote share after adjusting for depression rates and racial demographics. Each dot represents a state's deviation from the national average Republican vote percentage, with uncertainty intervals (95% credible intervals).

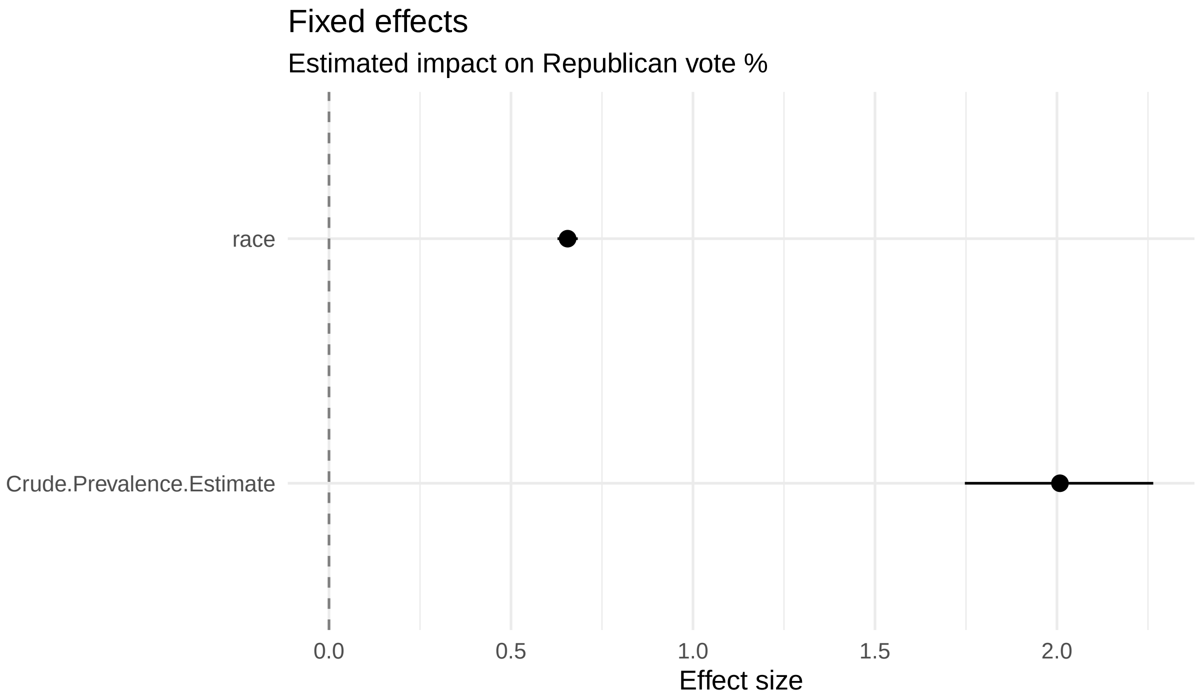

Figure 5: Estimated fixed effects from the hierarchical model. The plot shows the effect sizes of county-level depression rates and racial demographics on the Republican vote share, with uncertainty intervals representing the posterior estimates.

pattern (random effects). This addresses our objective's specific interest in "the effect that different states might have on the data result."

Finally, regular regression produced a residual error of 15.23, which dropped to just 9.24 with the hierarchical model, indicating a much better fit to our data. The large standard deviation between states (13.57) confirms that accounting for state-level variation was crucial for accurately understanding the depression-voting relationship.

Hierarchical modeling parameters

As for how I set up the model itself, I kept things straightforward but powerful. I included depression rates and race as fixed effects because our exploratory analysis showed these were the most relevant predictors. The correlation matrix revealed that population variables were less important for voting patterns, so I left those out to keep the model focused.

I decided to let each state have its own random intercept (the (1|STNAME) part in my model formula) because I wanted to capture how states might differ in their baseline Republican voting tendencies. Some states naturally lean more Republican or Democratic regardless of depression rates or demographics.

For the model fitting, I used the default priors in the brms package because they're reasonably uninformative for this type of data. I let the data speak for itself rather than imposing strong prior beliefs. The model ran for 2,000 iterations with 4 chains, giving plenty of samples to ensure our results are stable and reliable.

The results were fascinating - depression had an even stronger relationship with Republican voting (coefficient of 2.01) than our simpler models suggested, while properly accounting for the way voting patterns cluster by state. This modeling approach really helped us dig deeper into the relationship we were investigating.

Conclusion

Is there evidence that voting for Trump in the last election was associated with increased rates of depression?

After diving deep into the data from 3,107 counties across America, the answer to our question is a resounding yes—there is strong evidence that counties with higher depression rates showed stronger support for Trump in the 2024 election. But the story is far more intricate and fascinating than a simple yes or no.

When we look at the raw numbers, each 1% increase in depression prevalence is associated with approximately 2 percentage points higher Republican voting, even after accounting for racial demographics and state-level differences. This relationship is consistent and statistically significant. What's particularly striking is that this association became stronger when we properly accounted for state-level variations through hierarchical modeling.

This analysis revealed an America where mental health patterns and political preferences intertwine in complex ways. The state-level effects show dramatic geographic differences that persist even after controlling for depression and race—with Vermont counties voting about 40 percentage points more Democratic than predicted by demographics alone, while Alaska counties lean about 20 points more Republican. These findings suggest that voting patterns reflect not just demographic factors or economic conditions that are most frequently studied, but potentially also the collective psychological state of communities.

However, what's clear from our analysis is that understanding American political behavior requires looking beyond traditional factors to consider the role of mental health and well-being across communities. The significant state-level variations also remind us that America remains a patchwork of distinct regional political cultures, where local context continues to shape how voters express their preferences at the ballot box.

Appendices

Appendix A: Import required libraries

```
library(brms)
library(maps)
library(dplyr)
library(mapdata)
library(ggrepel)
library(ggplot2)
library(reshape2)
library(tidyverse)
library(ggcorrplot)
```

Appendix B: Generate summary of US votes data set

```
votes_data <- read.csv("USvotes.csv")
summary(votes_data)
```

```
##     STNAME            CTYNAME             TOT_POP           TOT_MALE
##  Length:3107        Length:3107        Min.   :     43   Min.   :     31
##  Class :character   Class :character   1st Qu.:  10922   1st Qu.:   5510
##  Mode  :character   Mode  :character   Median :  26125   Median :  13205
##                                        Mean   : 106488   Mean   :  52701
##                                        3rd Qu.:  69335   3rd Qu.:  34932
##                                        Max.   :9663345   Max.   :4780566
##   TOT_FEMALE       total_votes         per_gop       Crude.Prevalence.Estimate
##  Min.   :     12   Min.   :     97   Min.   : 5.081   Min.   :10.70
##  1st Qu.:   5374   1st Qu.:   5250   1st Qu.:58.260   1st Qu.:18.80
##  Median :  12980   Median :  12549   Median :70.274   Median :21.10
##  Mean   :  53787   Mean   :  49089   Mean   :66.947   Mean   :21.13
##  3rd Qu.:  34538   3rd Qu.:  33866   3rd Qu.:79.177   3rd Qu.:23.30
##  Max.   :4882779   Max.   :3728427   Max.   :95.965   Max.   :30.70
##       race
##  Min.   : 7.317
##  1st Qu.:79.183
##  Median :90.518
##  Mean   :83.948
##  3rd Qu.:94.760
##  Max.   :98.401
```

Appendix C: Plot average republican vote percentage by US states

```
state_summary <- votes_data %>%
  group_by(STNAME) %>%
  summarize(
    mean_gop = mean(per_gop),
    mean_depression = mean(Crude.Prevalence.Estimate),
```

```r
    mean_race = mean(race),
    n_counties = n()
  )

us_states <- map_data("state")
state_summary$state_lower <- tolower(state_summary$STNAME)

map_data <- left_join(us_states, state_summary, by = c("region" = "state_lower"))

ggplot(map_data, aes(x = long, y = lat, group = group, fill = mean_gop)) +
  geom_polygon(color = "white", linewidth = 0.2) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  scale_fill_gradient2(
    name = "Republican vote (%)",
    low = "blue", mid = "white", high = "red",
    midpoint = 50,
    limits = c(min(state_summary$mean_gop), max(state_summary$mean_gop))
  ) +
  labs(title = "Average republican vote percentage by state",
       subtitle = "2024 US Presidential Election") +
  theme_minimal() +
  theme(
    axis.text = element_blank(),
    axis.title = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank()
  )
```

Appendix D: Plot average depression rate by US states

```r
ggplot(map_data, aes(x = long, y = lat, group = group, fill = mean_depression)) +
  geom_polygon(color = "white", linewidth = 0.2) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  scale_fill_viridis_c(
    name = "Depression rate (%)",
    option = "plasma"
  ) +
  labs(title = "Average depression rate by state",
       subtitle = "Based on self-reported data") +
  theme_minimal() +
  theme(
    axis.text = element_blank(),
    axis.title = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank()
  )
```

Appendix E: Test correlation between voting, depression, race and demographics for US states

```
cor.test(votes_data$per_gop, votes_data$Crude.Prevalence.Estimate)
```

```
##
##  Pearson's product-moment correlation
##
## data:  votes_data$per_gop and votes_data$Crude.Prevalence.Estimate
## t = 14.674, df = 3105, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2214715 0.2872456
## sample estimates:
##      cor
## 0.254653
```

Appendix F: Plot correlation matrix for the data set

```
corr_matrix <- cor(votes_data[, c("per_gop", "Crude.Prevalence.Estimate", "race", "TOT_POP", "TOT_MALE"

melted_corr <- melt(corr_matrix)
names(melted_corr) <- c("Var1", "Var2", "value")

ggplot(melted_corr, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "navy", mid = "white", high = "firebrick",
                       midpoint = 0, limits = c(-1, 1), name = "Corr") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(title = "Correlation between voting, depression, race and demographics\nin recent US presidentia
       x = "", y = "") +
  geom_text(aes(label = round(value, 2)), size = 3)
```

Appendix G: Generate summary of linear regression on data set

```
model1 <- lm(per_gop ~ Crude.Prevalence.Estimate, data = votes_data)
summary(model1)
```

```
##
## Call:
## lm(formula = per_gop ~ Crude.Prevalence.Estimate, data = votes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.930  -8.482   2.795  10.520  35.161 
## 
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               39.92759    1.86152   21.45   <2e-16 ***
## Crude.Prevalence.Estimate  1.27861    0.08714   14.67   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 15.23 on 3105 degrees of freedom
## Multiple R-squared:  0.06485,    Adjusted R-squared:  0.06455 
## F-statistic: 215.3 on 1 and 3105 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(per_gop ~ Crude.Prevalence.Estimate + race, data = votes_data)
summary(model2)
```

```
## 
## Call:
## lm(formula = per_gop ~ Crude.Prevalence.Estimate + race, data = votes_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -51.208  -6.766   1.902   9.130  29.837 
## 
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                0.52193    1.91382   0.273    0.785    
## Crude.Prevalence.Estimate  1.07609    0.07352  14.636   <2e-16 ***
## race                       0.52038    0.01453  35.821   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 12.82 on 3104 degrees of freedom
## Multiple R-squared:  0.3384, Adjusted R-squared:  0.3379 
## F-statistic: 793.7 on 2 and 3104 DF,  p-value: < 2.2e-16
```

Appendix H: Generate summary of hierarchical multilevel modeling on data set

```
model_hier <- brms::brm(
  per_gop ~ Crude.Prevalence.Estimate + race + (1|STNAME),
  data = votes_data,
  family = gaussian()
)
```

```
summary(model_hier)
```

```
##  Family: gaussian 
##   Links: mu = identity; sigma = identity 
## Formula: per_gop ~ Crude.Prevalence.Estimate + race + (1 | STNAME) 
##    Data: votes_data (Number of observations: 3107) 
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Multilevel Hyperparameters:
## ~STNAME (Number of levels: 50)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    13.58      1.51    10.95    16.93 1.01      328      615
##
## Regression Coefficients:
##                           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
## Intercept                   -33.83      3.31   -40.27   -27.25 1.00      582
## Crude.Prevalence.Estimate     2.01      0.13     1.75     2.26 1.00     1194
## race                          0.66      0.01     0.63     0.68 1.00     2026
##                           Tail_ESS
## Intercept                     1031
## Crude.Prevalence.Estimate     1589
## race                          2675
##
## Further Distributional Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     9.24      0.12     9.02     9.47 1.00     2729     2475
##
## 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).
```

Appendix I: Compute the fixed effects

```
ranef_summary <- brms::ranef(model_hier, summary = TRUE)
state_effects_df <- data.frame(
  state = rownames(ranef_summary$STNAME),
  estimate = ranef_summary$STNAME[, "Estimate", "Intercept"],
  lower = ranef_summary$STNAME[, "Q2.5", "Intercept"],
  upper = ranef_summary$STNAME[, "Q97.5", "Intercept"]
)

state_effects_df <- state_effects_df[order(state_effects_df$estimate), ]
state_effects_df$state <- factor(state_effects_df$state, levels = state_effects_df$state)

# Get fixed effects
fixed_effects <- fixef(model_hier)
fixed_effects_df <- data.frame(
  term = rownames(fixed_effects),
  estimate = fixed_effects[, 1],
  lower = fixed_effects[, 3],  # this is 2.5% quantile
  upper = fixed_effects[, 4]   # this is 97.5% quantile
)
```

Appendix J: Plot the fixed effects

```
ggplot(fixed_effects_df[-1, ], aes(x = estimate, y = term)) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
  geom_pointrange(aes(xmin = lower, xmax = upper)) +
  labs(
    title = "Fixed effects",
    subtitle = "Estimated impact on republican vote %",
    x = "Effect size",
    y = ""
  ) +
  theme_minimal()
```

Appendix K: Plot state-effects with colors to highlight extreme states

```
ggplot(state_effects_df, aes(x = estimate, y = state, color = estimate > 0)) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
  geom_pointrange(aes(xmin = lower, xmax = upper)) +
  scale_color_manual(values = c("blue", "red"), guide = "none") +
  labs(
    title = "State-level effects on republican voting",
    subtitle = "After controlling for depression rates and racial demographics",
    x = "Effect on republican vote % (compared to national average)",
    y = "State"
  ) +
  theme_minimal()
```