Draft

STA 210 - Project

Bayes' Harem - Christina Wang, Kat Cottrell, David Goh, Ethan Song

```r
library(tidyverse)
library(tidymodels)
library(knitr)
library(ggfortify)
library(GGally)
```

```r
abortion_data_full<-read_csv(here::here("data/abortion-attitudes",
                                        "wvs-usa-abortion-attitudes-data.csv"))
```

Introduction and Research Question

Understanding public attitudes on divisive political issues is an important way for political leaders to mobilize voters and for lawmakers to draft laws that represent their constituents. While it may be easier to poll constituents' positions on an issue, it can be challenging to assess the complex factors that influence and predict those stances.

Abortion is one such divisive issue in the United States. Both pro-choice and pro-life groups have a history of mobilizing in states across the country in support of legislation for their respective sides (Ziegler, 2020). However, following the Supreme Court's 1973 decision in Roe v. Wade, a ruling which protected an individual's right to have an abortion before fetal viability, the issue has risen in political salience. Both the pro-choice and pro-life movements have gained national prominence, and the two major political parties have polarized around the issue, with the Democratic Party in favor of and the Republican Party against policies legalizing and increasing access to abortion (Weinberger 2022). Abortion has also increasingly become a key issue that voters consider when making their choice at the ballot box, with an increasing share of Americans identifying as "single-issue voters" regarding abortion (Brenan 2020).

In May 2022, a leaked draft opinion revealed that the US Supreme Court is prepared to overturn Roe v. Wade. Overturning Roe would dramatically change the trajectory of abortion politics in the US. Unless Congress passes a national policy, states would be able to decide whether or not to legalize abortion and gain much greater leverage in regulating access to the procedure (Weinberger 2022).

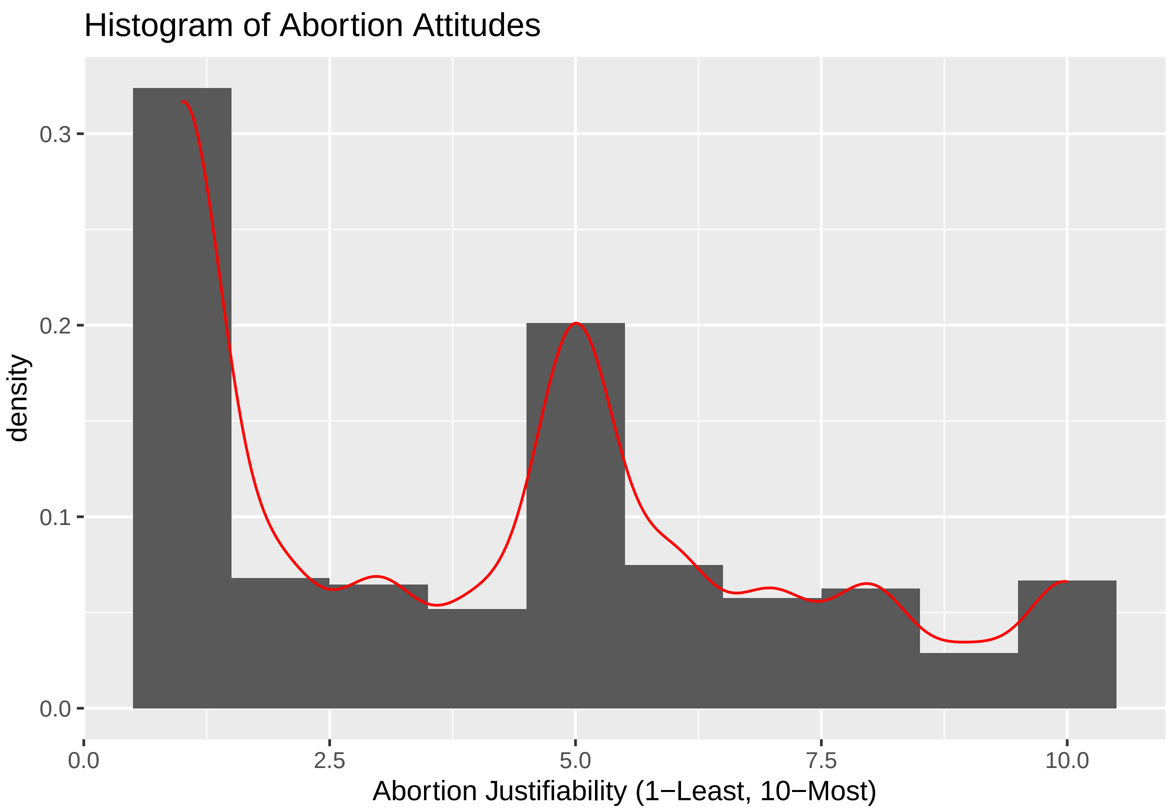
Given the potential overturning of Roe and the polarizing nature of the issue, it is important to understand how the American public feels about whether abortion should be legal or not, how accessible the procedure should be, and which factors influence these opinions. Understanding public opinion on the issue will ensure that political leaders are able to mobilize the correct constituencies, and that policy experts are able to pass policies on this issue that accurately reflect the preferences of the American people.

The dataset used in this analysis observes attitudes on the justifiability of abortion among respondents in in the United States across six “waves” of the World Values Survey (1982-2011), which is administered every few years and collects information about people’s values and beliefs worldwide, alongside basic demographic characteristics. The data are collected via face-to-face interviews at the respondents’ homes. In this data set, the response variable, justifiability of abortion, is a numerical measure on a scale of 1 to 10 on the individual person’s attitude toward whether abortion is justifiable or not. Individuals responded 1 for “abortion is never justified” and 10 for “abortion is always justified.

Our research question is as follows: “Do an individual’s political ideology and their personal attitudes/preferences towards other issues, such as the importance of religion in their life and their respect for authority, among others, predict their attitude on the justifiability of abortion?” We will attempt to answer this question using an EDA-informed predictive model. Given that this issue has become highly polarized by political party, we predict that liberal ideology and liberal-leaning attitudes on other issues will correlate with belief that abortion is more justified.

Exploratory Data Analysis

```
ggplot(abortion_data_full, aes(x = aj)) +
  geom_histogram(binwidth = 1, aes(y=..density..)) +
  geom_density(color = "red") +
  labs(title = "Histogram of Abortion Attitudes",
       x = "Abortion Justifiability (1-Least, 10-Most)")
```



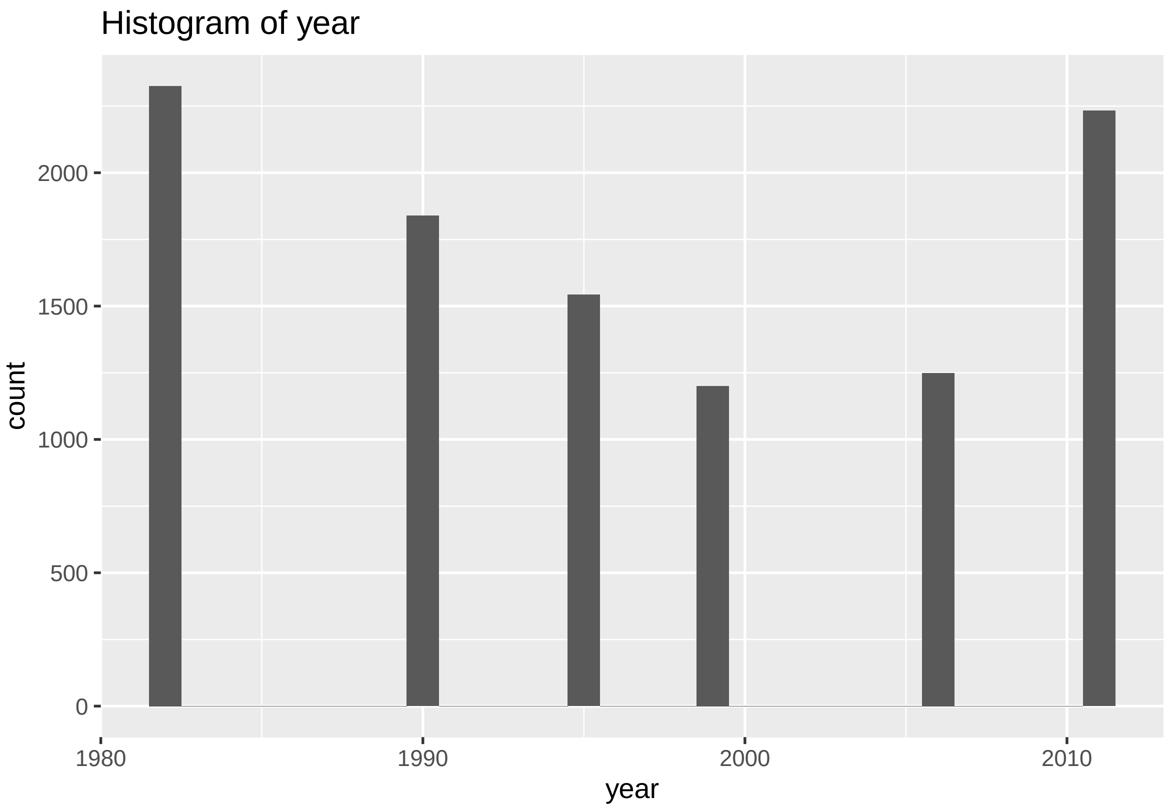
```
summary(abortion_data_full$aj)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  1.000   1.000   4.000   4.147   6.000  10.000     299 
```

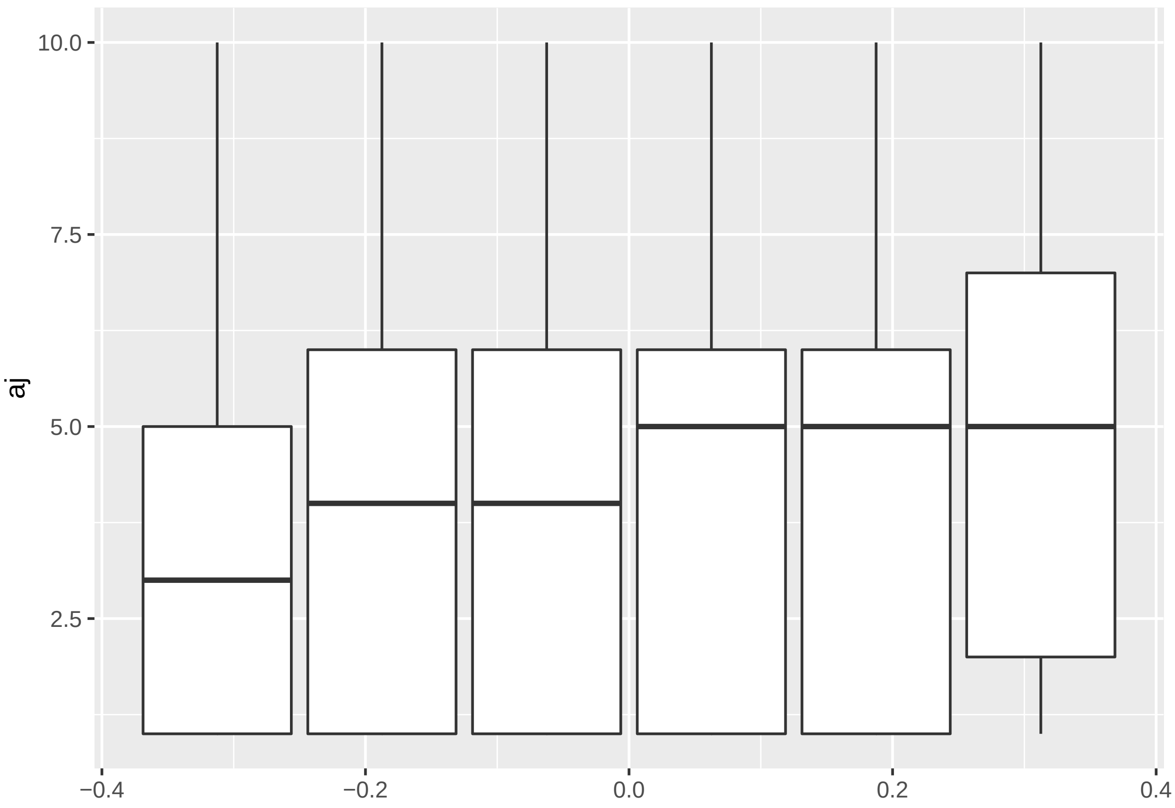
From the above visualization of the distribution of attitudes toward the justifiability of abortion, we observe that it is not a bell shape, but rather trimodal. This is likely because the question's phrasing is similar to a yes/no question, but respondents were asked to give their level of agreement on a scale of 1-10. This may worsen the ability of a multiple linear regression (MLR) model to fit the data. Consequently, further research may choose to truncate the attitude on the justifiability of abortion into a categorical variable such as (Agree, Disagree, Undecided), conducting a binomial or multinomial logistic regression thereafter.

We now extend our exploratory data analysis (EDA) to the predictor variables of interest: year of survey, ideology, Child Autonomy Index, Importance of God, Respect for Authority, and National Pride. The EDA for each variable comprises a histogram and a boxplot of the response variable, grouped by value of the predictor. We also provide a correlation matrix to detect any multicollinearity between our predictor variables, which would increase the uncertainty of our model's parameters.

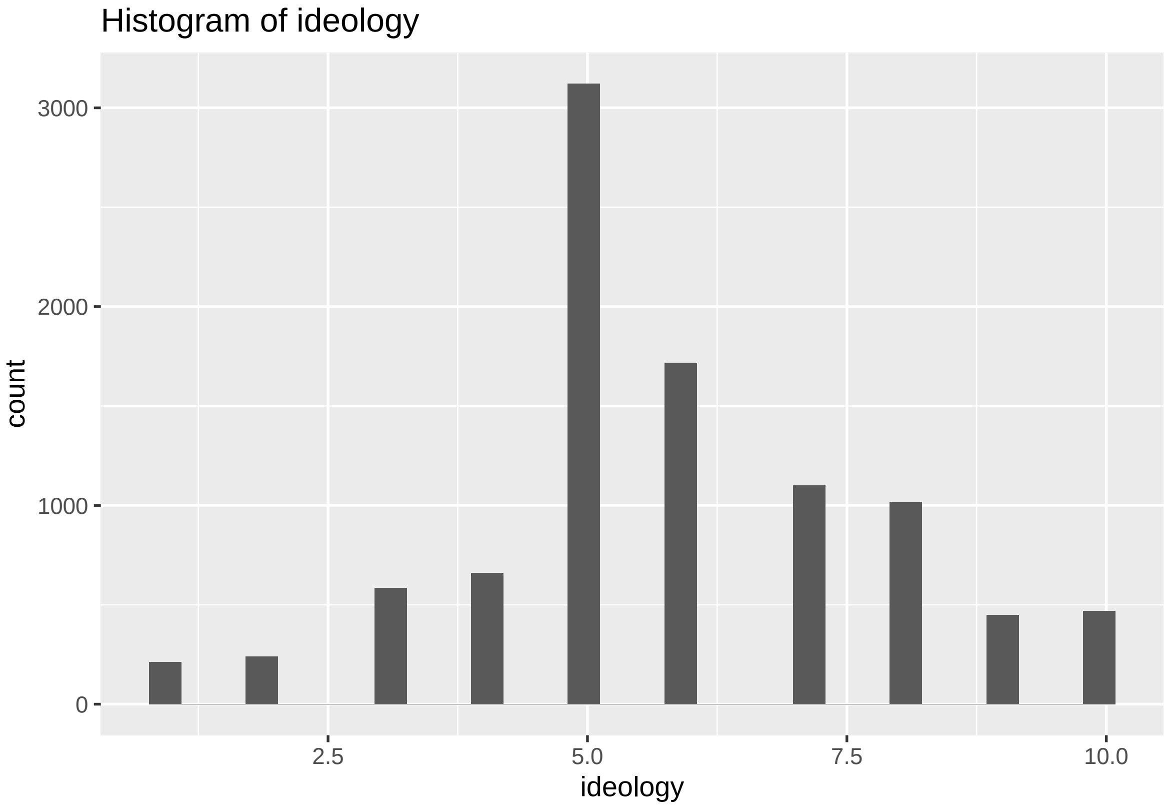
```r
#Year of survey
ggplot(abortion_data_full, aes(x = year)) +
  geom_histogram() +
  labs(title = "Histogram of year")
```



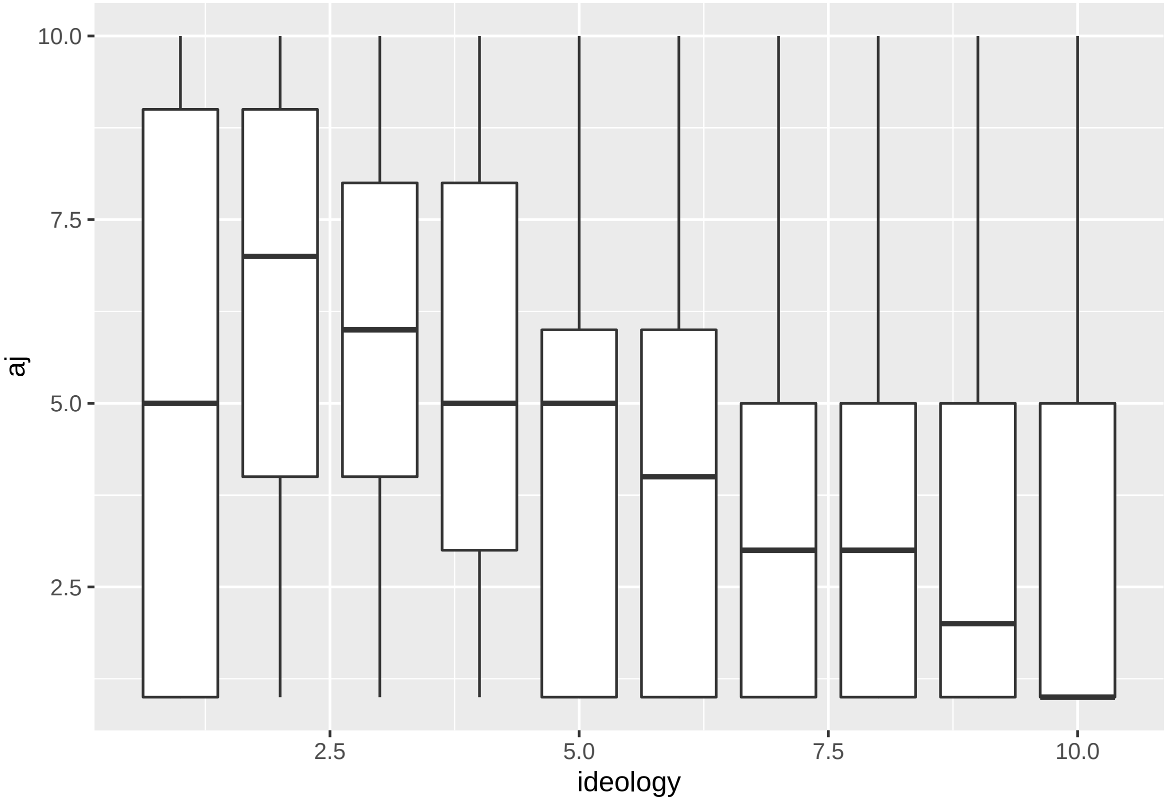
```r
ggplot(abortion_data_full, aes(group = year, y = aj)) +
  geom_boxplot()
```



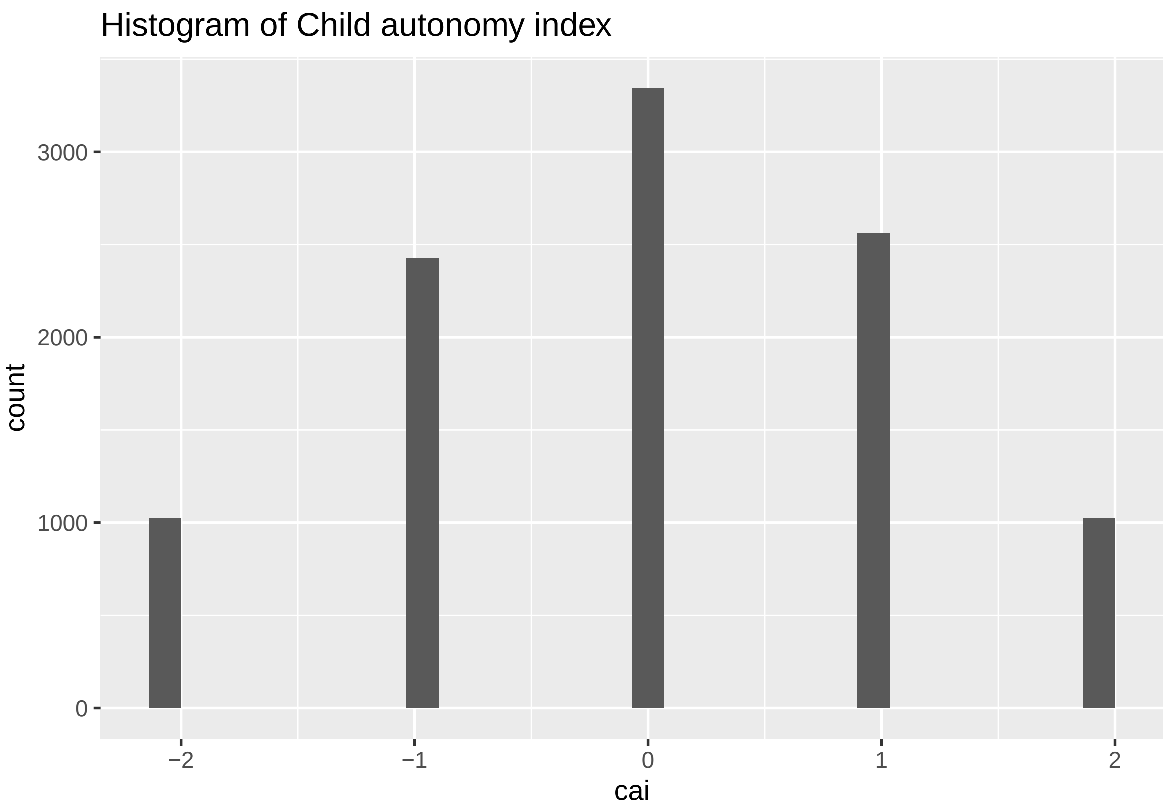
```
# ideology
ggplot(abortion_data_full, aes(x = ideology)) +
  geom_histogram() +
  labs(title = "Histogram of ideology")
```



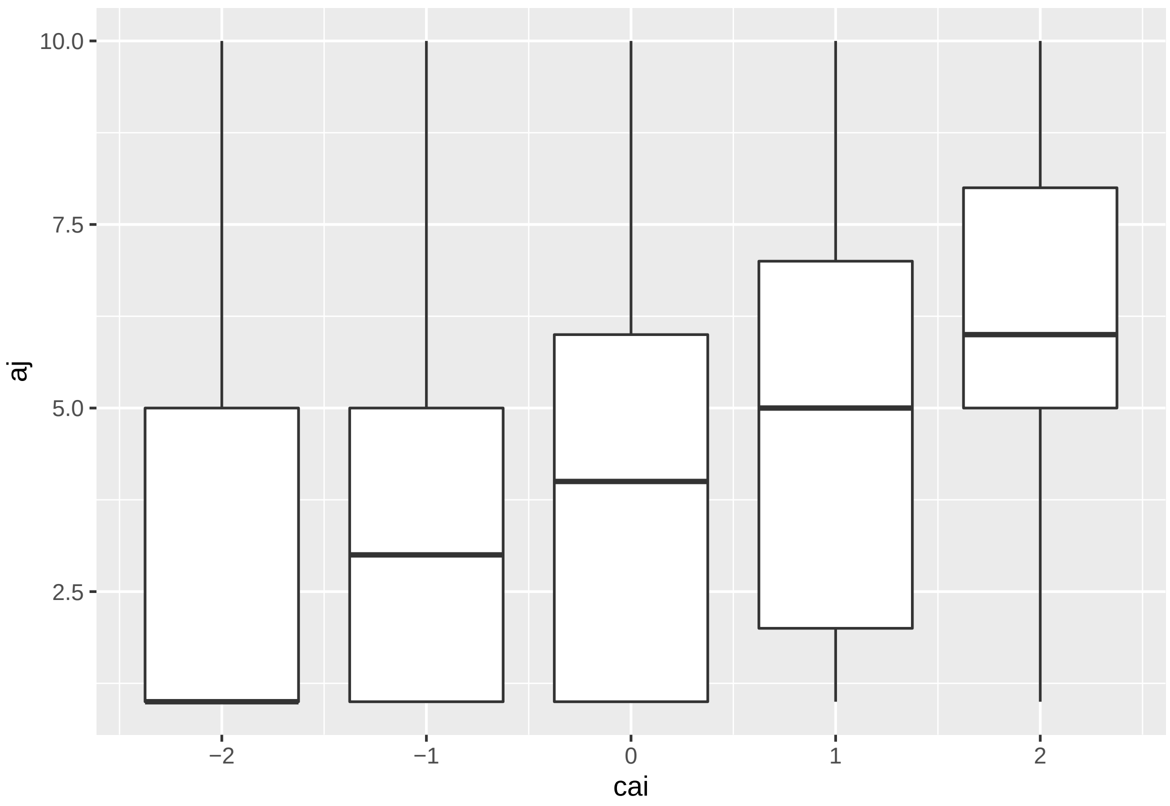
```
ggplot(abortion_data_full, aes(x = ideology,group = ideology, y = aj)) +
  geom_boxplot()
```



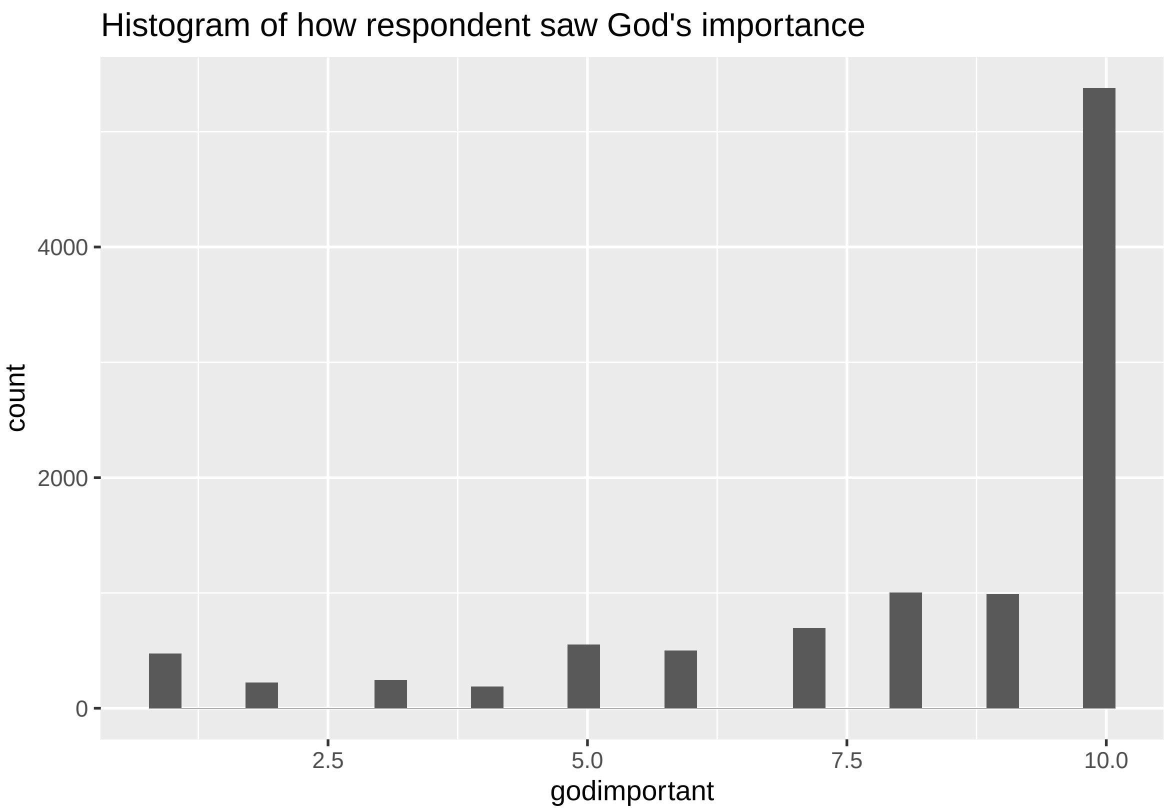
```
# Child autonomy index
ggplot(abortion_data_full, aes(x = cai)) +
  geom_histogram() +
  labs(title = "Histogram of Child autonomy index")
```



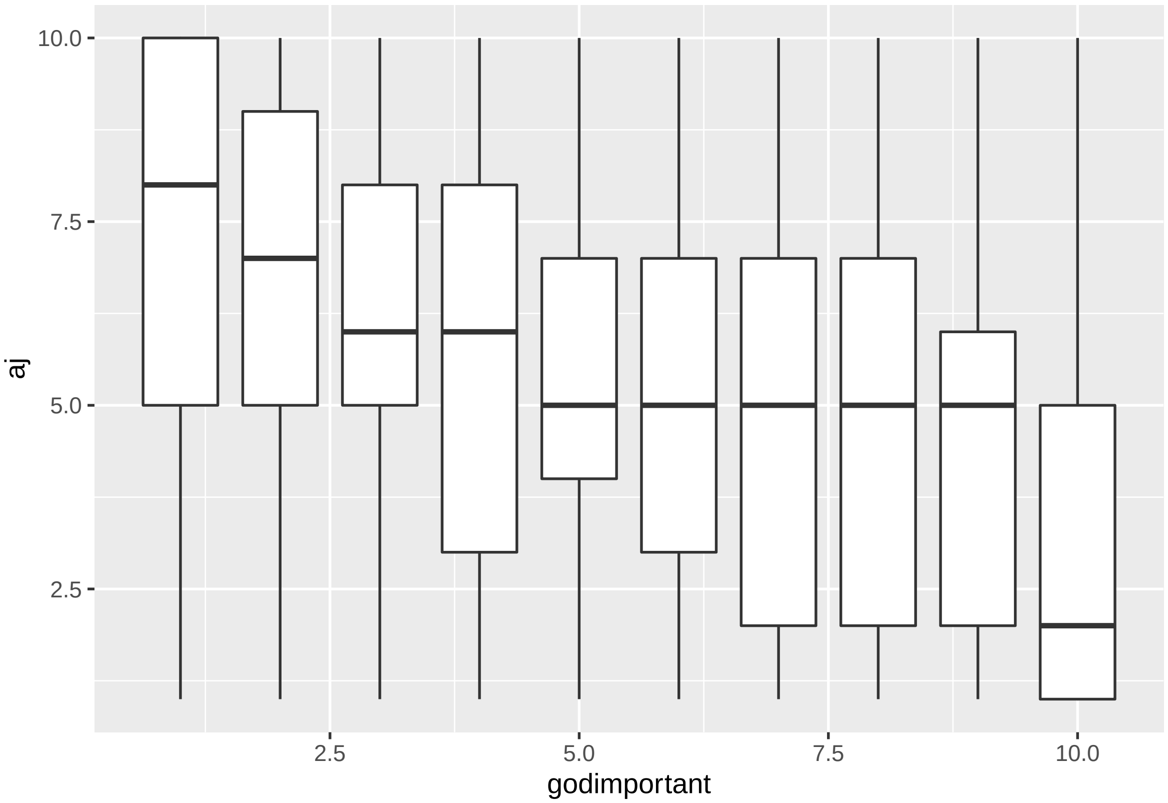
```
ggplot(abortion_data_full, aes(x = cai, group = cai, y = aj)) +
  geom_boxplot()
```



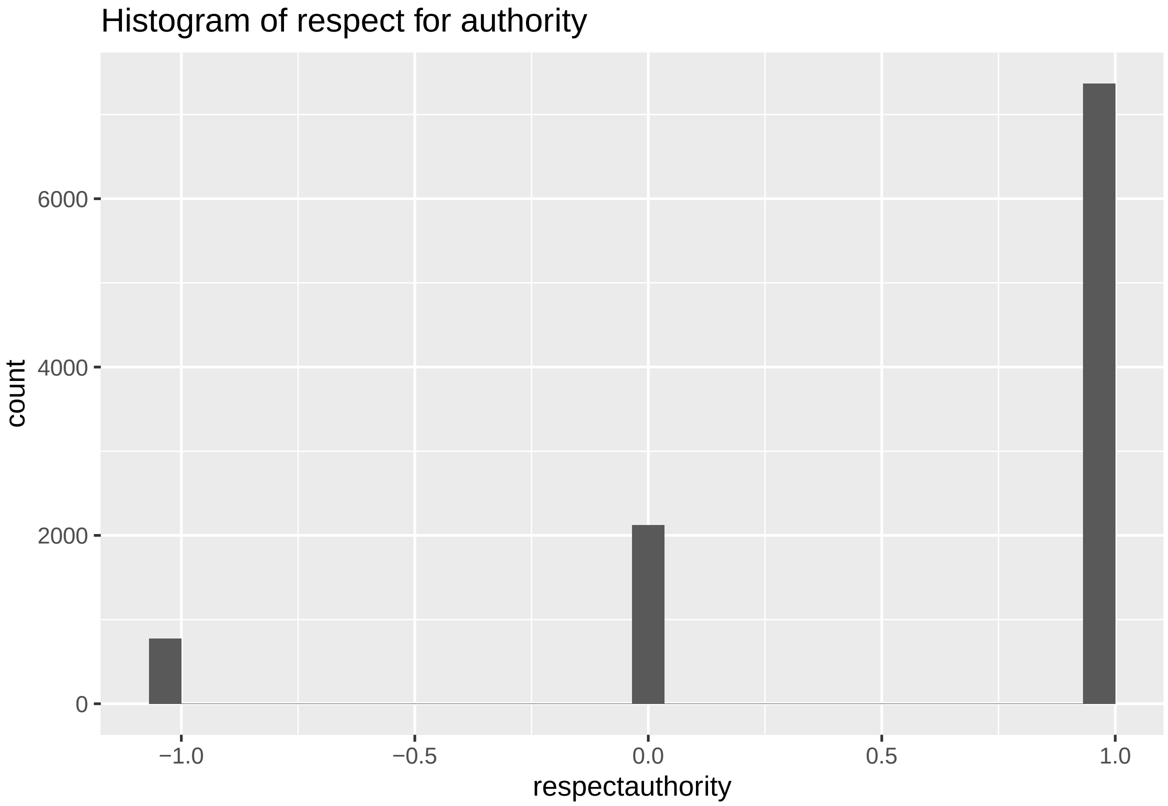
```
# Importance of God
ggplot(abortion_data_full, aes(x = godimportant)) +
  geom_histogram() +
  labs(title = "Histogram of how respondent saw God's importance")
```



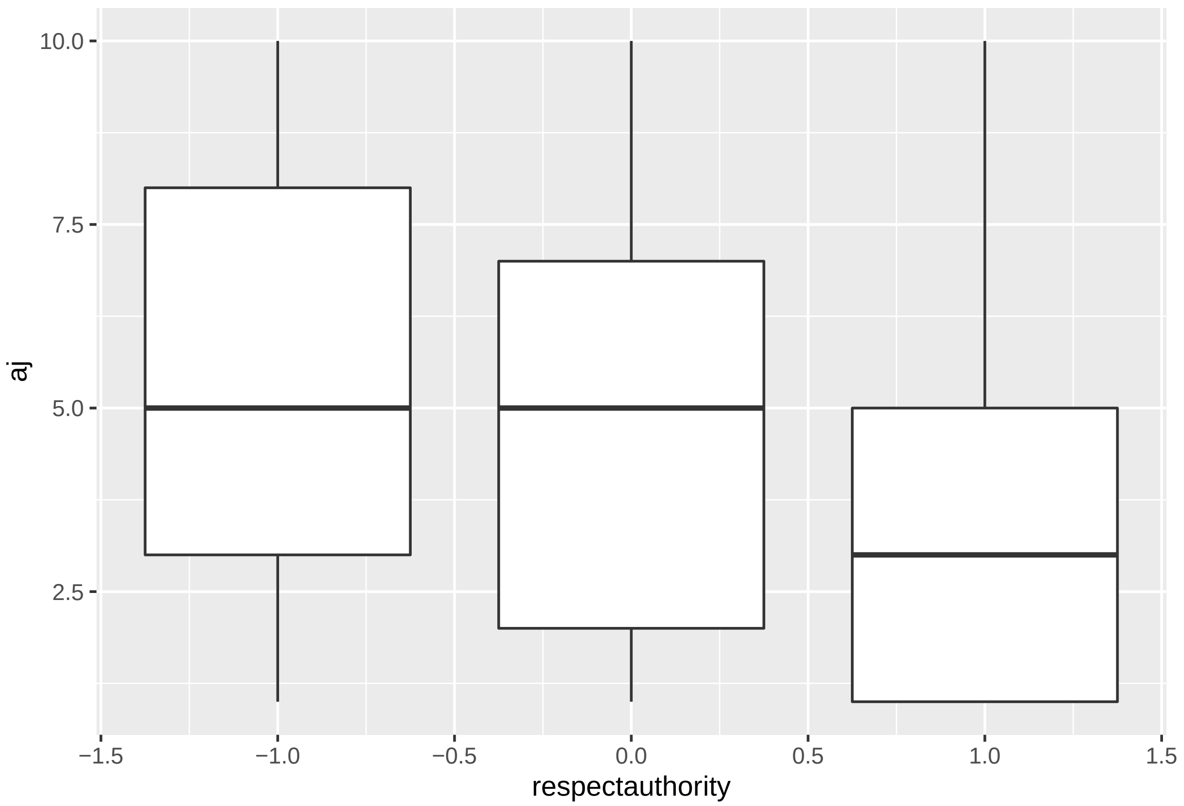
```
ggplot(abortion_data_full, aes(x = godimportant, group = godimportant,
                               y = aj)) +
  geom_boxplot()
```



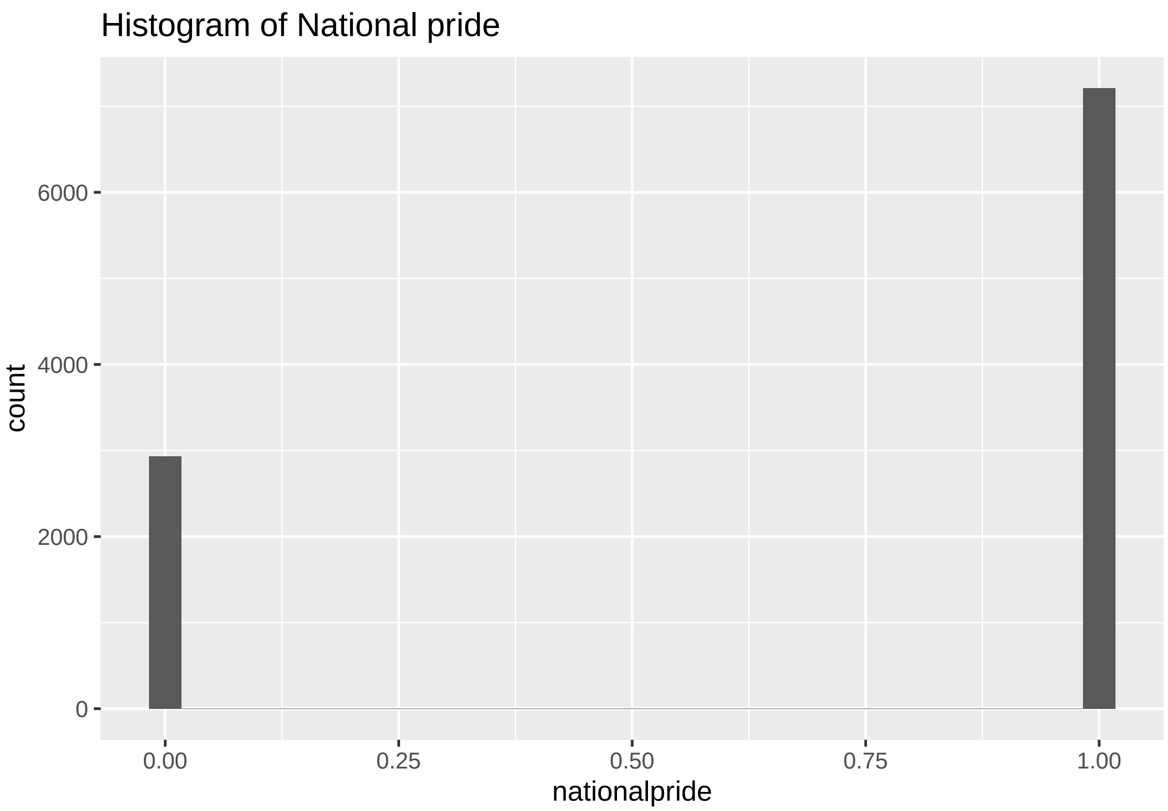
```r
# Respect for authority
ggplot(abortion_data_full, aes(x = respectauthority)) +
  geom_histogram() +
  labs(title = "Histogram of respect for authority")
```



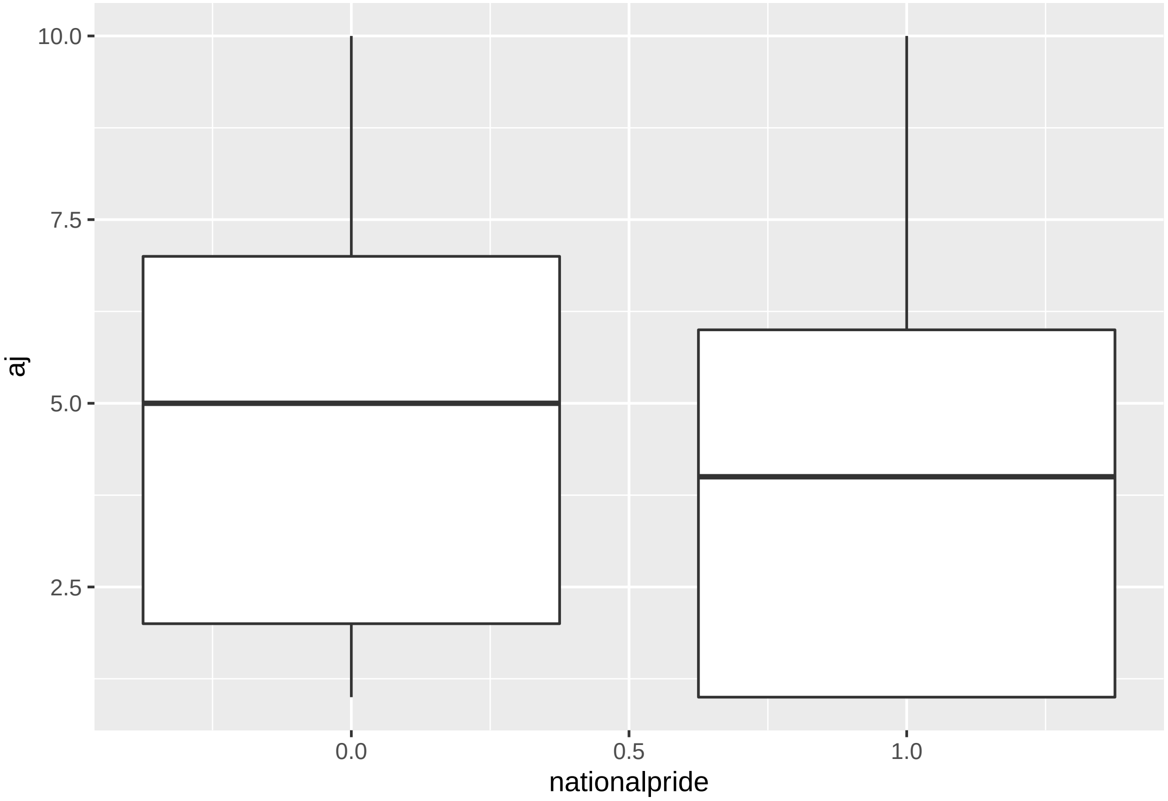
```
ggplot(abortion_data_full, aes(x = respectauthority, group = respectauthority,
                               y = aj)) +
  geom_boxplot()
```



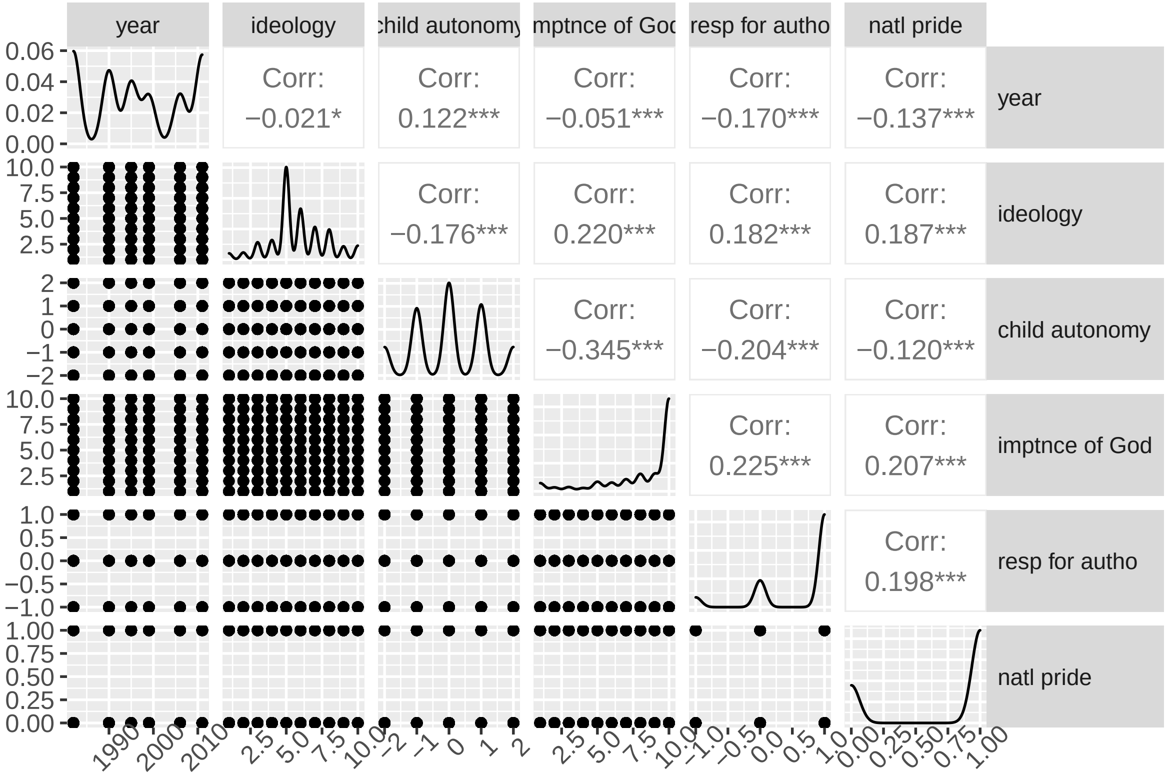
```
# National Pride
ggplot(abortion_data_full, aes(x = nationalpride)) +
  geom_histogram() +
  labs(title = "Histogram of National pride")
```



```
ggplot(abortion_data_full, aes(x = nationalpride, group = nationalpride, y = aj)) +
  geom_boxplot()
```



```
ggpairs(abortion_data_full,
  columns = c("year", "ideology", "cai", "godimportant", "respectauthority",
              "nationalpride"),
  columnLabels = c("year", "ideology", "child autonomy", "imptnce of God",
                   "resp for autho", "natl pride")) +
   theme(
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(angle = 45, size = 10),
    strip.text.y = element_text(angle = 0, hjust = 0)
    )
```


year
ideology
child autonomy
imptnce of God
resp for autho
natl pride
0.06
0.04
0.02
0.00
Corr:
−0.021*
Corr:
0.122***
Corr:
−0.051***
Corr:
−0.170***
Corr:
−0.137***
10.0
7.5
5.0
2.5
Corr:
−0.176***
Corr:
0.220***
Corr:
0.182***
Corr:
0.187***
2
1
0
−1
−2
Corr:
−0.345***
Corr:
−0.204***
Corr:
−0.120***
Corr:
0.225***
Corr:
0.207***
1.0
0.5
0.0
−0.5
−1.0
Corr:
0.198***
1.00
0.75
0.50
0.25
0.00
1990
2000
2010
2.5
5.0
7.5
10.0
−2
−1
0
1
2
−1.0
−0.5
0.0
0.5
1.0
0.00
0.25
0.50
0.75
1.00

Methodology

Given that our response variable, attitude towards the justifiability of abortion, is measured on a numeric scale from 1 to 10, and that there are multiple predictor variables that we suspect to be potentially related to it, we will conduct multiple linear regression (MLR) to model the effect of these predictors on the variation in abortion attitude.

We chose to consider predictor variables that, from our exploratory data analysis (EDA), looked like they were potentially correlated with the outcome. Because the predictor variables were discrete, our EDA was presented as a series of box plots. We observed the differences in median and quartile values of abortion attitude across different predictor values, ruling out variables with no visible hint of relationship. This led to the following predictor variables: Year of survey, Political Ideology, Child Autonomy Index, Importance of God, Respect for Authority, and National Pride.

A brief note on data cleaning: many survey questions were only added from 1995 onward, and observations before 1995 have a high number of missing values; therefore, we removed the observations that were part of "generational waves" before 1995.

```
abortion_data <- abortion_data_full %>%
  filter(year >= "1995") %>%
  select(-starts_with("wave"), -starts_with("wvsccode"))
```

Fitting the Model

We conducted a 75%-25% data split into training and testing sets, using the seed "206". Fitting the MLR, we obtain the following table of parameter point estimates.

```
set.seed(206)
abortion_split <- initial_split(abortion_data)
abortion_training <- training(abortion_split)
abortion_testing <- testing(abortion_split)
```

```
abortion_spec <- linear_reg() %>%
  set_engine("lm")
```

```
abortion_rec1 <- recipe(aj ~ year + ideology + cai + godimportant +
                        respectauthority, nationalpride,
                        data = abortion_data) %>%
  step_center(year) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors())
```

```
abortion_rec1
```

```
Recipe

Inputs:

      role #variables
   outcome          1
 predictor          5

Operations:

Centering for year
Dummy variables from all_nominal_predictors()
Zero variance filter on all_predictors()
```

```
abortion_rec2 <- recipe(aj ~ year + godimportant,
                        data = abortion_data) %>%
  step_center(year) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors())

abortion_rec2
```

```
Recipe

Inputs:

      role #variables
   outcome          1
 predictor          2

Operations:

Centering for year
Dummy variables from all_nominal_predictors()
Zero variance filter on all_predictors()
```

```
abortion_rec3 <- recipe(aj ~ year + ideology + cai + nationalpride,
                        data = abortion_data) %>%
  step_center(year) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors())

abortion_rec3
```

```
Recipe

Inputs:

      role #variables
   outcome          1
 predictor          4

Operations:

Centering for year
Dummy variables from all_nominal_predictors()
Zero variance filter on all_predictors()
```

```
abortion_wflow1 <- workflow() %>%
  add_model(abortion_spec) %>%
  add_recipe(abortion_rec1)

abortion_wflow2 <- workflow() %>%
  add_model(abortion_spec) %>%
  add_recipe(abortion_rec2)

abortion_wflow3 <- workflow() %>%
  add_model(abortion_spec) %>%
  add_recipe(abortion_rec3)
```

```
abortion_fit1 <- abortion_wflow1 %>%
  fit(data = abortion_training)
tidy(abortion_fit1) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	8.500	0.162	52.393	0.000
year	0.029	0.006	4.754	0.000

term	estimate	std.error	statistic	p.value
ideology	-0.284	0.021	-13.741	0.000
cai	0.469	0.037	12.766	0.000
godimportant	-0.288	0.016	-18.224	0.000
respectauthority	-0.190	0.068	-2.795	0.005

```r
abortion_fit2 <- abortion_wflow2 %>%
  fit(data = abortion_training)
tidy(abortion_fit2) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	7.914	0.125	63.183	0
year	0.035	0.006	5.782	0
godimportant	-0.429	0.015	-29.349	0

```r
abortion_fit3 <- abortion_wflow3 %>%
  fit(data = abortion_training)
tidy(abortion_fit3) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	6.653	0.133	49.865	0
year	0.034	0.006	5.466	0
ideology	-0.348	0.021	-16.386	0
cai	0.704	0.036	19.641	0
nationalpride	-0.402	0.090	-4.488	0

[table]

Hence, our MLR model equation is as follows, where x1 is **, x2 is** , x3 is **, x4 is** , and x5 is ___:

[equation in latex]

Model Inference and Prediction

Inference: CI and HT

We will take a quick look at the 95% confidence intervals of the parameters from our fitted model, displayed in the following plots:

[plots of parameters' .estimate against the parameter names]

```
tidy(abortion_fit1, conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	8.500	0.162	52.393	0.000	8.182	8.818
year	0.029	0.006	4.754	0.000	0.017	0.041
ideology	-0.284	0.021	-13.741	0.000	-0.325	-0.244
cai	0.469	0.037	12.766	0.000	0.397	0.541
godimportant	-0.288	0.016	-18.224	0.000	-0.319	-0.257
respectauthority	-0.190	0.068	-2.795	0.005	-0.323	-0.057

```
tidy(abortion_fit2, conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.914	0.125	63.183	0	7.669	8.160
year	0.035	0.006	5.782	0	0.023	0.047
godimportant	-0.429	0.015	-29.349	0	-0.457	-0.400

```
tidy(abortion_fit3, conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	6.653	0.133	49.865	0	6.391	6.914
year	0.034	0.006	5.466	0	0.022	0.047
ideology	-0.348	0.021	-16.386	0	-0.390	-0.306
cai	0.704	0.036	19.641	0	0.634	0.775
nationalpride	-0.402	0.090	-4.488	0	-0.578	-0.227

We then proceed to our hypothesis test. As stated in our introduction, given that this issue has become highly polarized by political party, we suspect that liberal political attitudes will correlate with belief that abortion is more justified, leading to the following hypotheses.

[write out null hypothesis and alternative hypothesis]

From the p-value in our output displayed in Figure 5(?), we conclude …

Prediction:

We now consider the strength of our model in predicting the attitude toward the justifiability of abortion of a person of certain characteristics.

For a person who [describe liberal characteristics here], our model predicts that their attitude toward the justifiability of abortion is [write what the model would predict].

This does/does not support our hypothesis based on contextual knowledge of this issue.

However, calculating the R-squared and root-mean-squared-error (RMSE) values of our model shows that our model is not very strong at predicting attitude toward the justifiability of abortion, with a R^2 of ___ and RMSE of ___:

[R^2, RMSE output]

```r
abortion_training_pred1 <- predict(abortion_fit1, abortion_training) %>%
  bind_cols(abortion_training %>% select(aj, year, ideology, cai,
                                         godimportant, respectauthority,
                                         nationalpride))
abortion_training_pred1
```

```
# A tibble: 4,667 x 8
   .pred    aj  year ideology   cai godimportant respectauthority nationalpride
   <dbl> <dbl> <dbl>    <dbl> <dbl>        <dbl>            <dbl>         <dbl>
 1  5.46     9  2011        9     2            5                1             1
 2  5.01     5  1999        4     1           10               -1             1
 3 NA        3  2006       NA     0            9                1             0
 4  6.03     5  2011        5     1            6                0             0
 5  3.87     3  1995        6    -1            7                1             1
 6  4.41     1  2011        5     0           10                0             1
 7  3.28     1  2011        5    -2           10                1             1
 8  5.56     5  2011        4     0            7                0             1
 9  6.36     1  1999        5     1            3                1             0
10  6.15    10  1995        5     1            4                0             1
# ... with 4,657 more rows
```

```r
abortion_training_pred2 <- predict(abortion_fit2, abortion_training) %>%
  bind_cols(abortion_training %>% select(aj, year, godimportant))
abortion_training_pred2
```

```
# A tibble: 4,667 x 4
   .pred    aj  year godimportant
   <dbl> <dbl> <dbl>        <dbl>
 1  6.03     9  2011            5
 2  3.46     5  1999           10
 3  4.14     3  2006            9
 4  5.60     5  2011            6
 5  4.61     3  1995            7
 6  3.88     1  2011           10
 7  3.88     1  2011           10
 8  5.17     5  2011            7
 9  6.46     1  1999            3
10  5.89    10  1995            4
# ... with 4,657 more rows
```

```
abortion_training_pred3 <- predict(abortion_fit3, abortion_training) %>%
  bind_cols(abortion_training %>% select(aj, year, ideology, cai,
                                         nationalpride))
abortion_training_pred3
```

```
# A tibble: 4,667 x 6
   .pred    aj  year ideology   cai nationalpride
   <dbl> <dbl> <dbl>    <dbl> <dbl>         <dbl>
 1  4.78     9  2011        9     2             1
 2  5.40     5  1999        4     1             1
 3 NA        3  2006       NA     0             0
 4  5.87     5  2011        5     1             0
 5  3.16     3  1995        6    -1             1
 6  4.76     1  2011        5     0             1
 7  3.35     1  2011        5    -2             1
 8  5.11     5  2011        4     0             1
 9  5.46     1  1999        5     1             0
10  4.92    10  1995        5     1             1
# ... with 4,657 more rows
```

We will discuss potential reasons for this lackluster performance in our model diagnosis and conclusion sections; however, we will first conduct model evaluation and comparison to ascertain if our chosen model is still a better fit to the data compared to some alternatives.

A quick glance at our data might seem that the models are not that strong at predicting the abortion attitude (aj) of a given observation. We will look into this further now with cross-validation, then a test in AIC and BIC statistics.

Model Evaluation

To see how well our model fits the data on the attitudes toward the justifiability of abortion, we conducted a 10-fold cross-validation, calculating various metrics – the Akaike information criterion (AIC), Bayesian Information Criterion (BIC) and adjusted R-squared value – to evaluate the performance of our model on 10 subsets of our data.

We also compared this to two alternative models that used other combinations of predictors from the data set.

[3 line plots, displaying AIC, BIC and R^2 value respectively, against Fold number. Each plot should have the line graph for each of the 3 models we used)

```r
set.seed(206)
folds <- vfold_cv(abortion_training, v = 10)
abortion_fit_rs1 <- abortion_wflow1 %>%
  fit_resamples(folds)

abortion_fit_rs2 <- abortion_wflow2 %>%
  fit_resamples(folds)

abortion_fit_rs3 <- abortion_wflow3 %>%
  fit_resamples(folds)
```

```r
cv_metrics1 <- collect_metrics(abortion_fit_rs1, summarize = FALSE)
cv_metrics2 <- collect_metrics(abortion_fit_rs2, summarize = FALSE)
cv_metrics3 <- collect_metrics(abortion_fit_rs3, summarize = FALSE)

cv_metrics1 %>%
  mutate(.estimate = round(.estimate, 3)) %>%
  pivot_wider(id_cols = id, names_from = .metric, values_from = .estimate) %>%
  kable(col.names = c("Fold", "RMSE", "R-squared"))
```

Fold	RMSE	R-squared
Fold01	2.606	0.213
Fold02	2.490	0.236
Fold03	2.575	0.263
Fold04	2.694	0.182
Fold05	2.428	0.274
Fold06	2.526	0.269
Fold07	2.500	0.288
Fold08	2.513	0.280

Fold	RMSE	R-squared
Fold09	2.391	0.261
Fold10	2.579	0.196

```
cv_metrics2 %>%
  mutate(.estimate = round(.estimate, 3)) %>%
  pivot_wider(id_cols = id, names_from = .metric, values_from = .estimate) %>%
  kable(col.names = c("Fold", "RMSE", "R-squared"))
```

Fold	RMSE	R-squared
Fold01	2.674	0.181
Fold02	2.634	0.146
Fold03	2.667	0.202
Fold04	2.757	0.139
Fold05	2.591	0.192
Fold06	2.687	0.179
Fold07	2.640	0.191
Fold08	2.611	0.227
Fold09	2.573	0.155
Fold10	2.705	0.115

```
cv_metrics3 %>%
  mutate(.estimate = round(.estimate, 3)) %>%
  pivot_wider(id_cols = id, names_from = .metric, values_from = .estimate) %>%
  kable(col.names = c("Fold", "RMSE", "R-squared"))
```

Fold	RMSE	R-squared
Fold01	2.739	0.132
Fold02	2.579	0.182
Fold03	2.681	0.197
Fold04	2.748	0.142
Fold05	2.510	0.230
Fold06	2.629	0.207
Fold07	2.572	0.250
Fold08	2.708	0.159
Fold09	2.483	0.202
Fold10	2.630	0.164

```
glance(abortion_fit1) %>%
  select(AIC, BIC)
```

```
# A tibble: 1 x 2
     AIC    BIC
   <dbl>  <dbl>
1 19980. 20025.
```

```
glance(abortion_fit2) %>%
  select(AIC, BIC)
```

```
# A tibble: 1 x 2
     AIC    BIC
   <dbl>  <dbl>
1 21492. 21517.
```

```
glance(abortion_fit3) %>%
  select(AIC, BIC)
```

```
# A tibble: 1 x 2
     AIC    BIC
   <dbl>  <dbl>
1 20330. 20369.
```

From the above AIC and BIC values, we have evidence that our chosen model is the strongest model and best fit for the data compared to two alternatives with fewer predictors. Our model has the lowest AIC and BIC values, which are usually penalized for having more predictors, but in this case the model is strong enough to overcome that penalty.

Model Diagnosis

Conditions check

Linearity [residual plot vs fitted values]

Randomness

Independence [residual plot vs id]

Conclusion + Discussion

[paste here after finalizing the google doc]

Data dictionary

The data dictionary can be found here.

References

Brenan, M. (2020, July 7). One in Four Americans Consider Abortion a Key Voting Issue. Gallup. https://news.gallup.com/poll/313316/one-four-americans-consider-abortion-key-voting-issue.aspx

Weinberger, J. (2022, May 6). How we got here: Roe v. Wade from 1973 to today. Vox. https://www.vox.com/23055389/roe-v-wade-timeline-abortion-overturn-political-polarization

Ziegler, M. (2020, October 22). Abortion politics polarized before Roe. When it's gone, the fighting won't stop. The Washington Post. https://www.washingtonpost.com/outlook/2020/10/22/roe-polarize-abortion-politics/