# Proposal

## STA 210 - Project

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

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

```
abortion_data_full <- read_csv(here::here("data/abortion-attitudes", "wvs-usa-abortion-attitu
```

### Introduction

Our project is on the topic of abortion attitudes in the US. Abortion has long been a divisive issue in the US, with both pro-choice and pro-life groups organizing in states across the country in support of political change that supported their respective sides (Ziegler, 2020). However, following the Supreme Court's decision in Roe v. Wade in 1973, 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 they decide on candidates to vote for, with an increasing share of Americans becoming "single-issue voters" regarding abortion (Brenan 2020).

In May 2022, a draft opinion was leaked revealing that the Supreme Court is prepared to overturn Roe v. Wade. If Roe is overturned, then it 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 would 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 what factors influence these opinions. Understanding public opinion on the issue will ensure that political leaders will be 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.

In this project, we are researching the factors that influence Americans' abortion attitudes. Specifically, we would like to know which demographic factors (such as age, gender, and education level) and personal attitudes towards other issues (such as political ideology, importance of religion, and respect for authority) best predict an individual's attitude towards abortion. We will conduct our analysis using data collected about Americans' abortion attitudes, demographic information, as well as attitudes about other issues. The data are a nationally representative sample of the American people; hence we can infer that all the observations are independent and the variables may show a linear relationship. Given that this issue has become highly polarized by political party, we predict that holding liberal political ideology and liberal attitudes on other issues, as well as being younger in age is correlated with believing that abortion is more justified. (the exact predictors we will use will be decided after model selection).

Works Cited

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/

## Data description

The dataset observes "attitudes on the justifiability of abortion in the United States across six waves of World Values Survey data" (README.md) and some basic qualities of the respondants.

Observations include:

- WVS country code
- Generational wave (1982, 1990, 1995, 1999, 2006, or 2011)
- Justifiability of abortion (1-10)
- Age (17 to 96)
- College graduate (1 for yes)

- Female (1 for women) - Unemployed (1 = currently unemployed)
- Ideology (1-10 for left-right)
- Financial satisfaction (1-10 for least-most)
- WVS post-materialist index (-1 = materialist. 2 = mixed. 3 = post-materialist)
- Child autonomy index (-2 to 2 for obedience and religious faith-determination and independence)
- Trust (1 = believes most people can be trusted)
- Importance of God (1-10)
- Opinion of respect for authority (-1-1 for bad to good)
- National pride (1 = very proud to be an American)

The data were collected as part of the World Values Survey, which is administered every few years and collects information about people's values and beliefs worldwide. The survey aims to get a nationally representative sample of a minimum of 1200 for most countries, and the data are collected via face-to-face interviews at the respondents' homes. The data included in this set specifically include responses from 6 waves of the survey (administered over the period 1982-2011). The responses included in this set are from people in the United States, and it examines their attitudes towards abortion.

## Analysis approach

The response variable, **"Justifiability of abortion"**, is a numerical measure on a scale of 1-10 on the individual person's attitude toward whether abortion is justifiable or not. The individuals responded "1" for "abortion is never justified" and "10" for "abortion is always justified.

We will conduct a **Multiple Linear Regression (MLR)** on this response variable against 6 other predictor variables for our project.

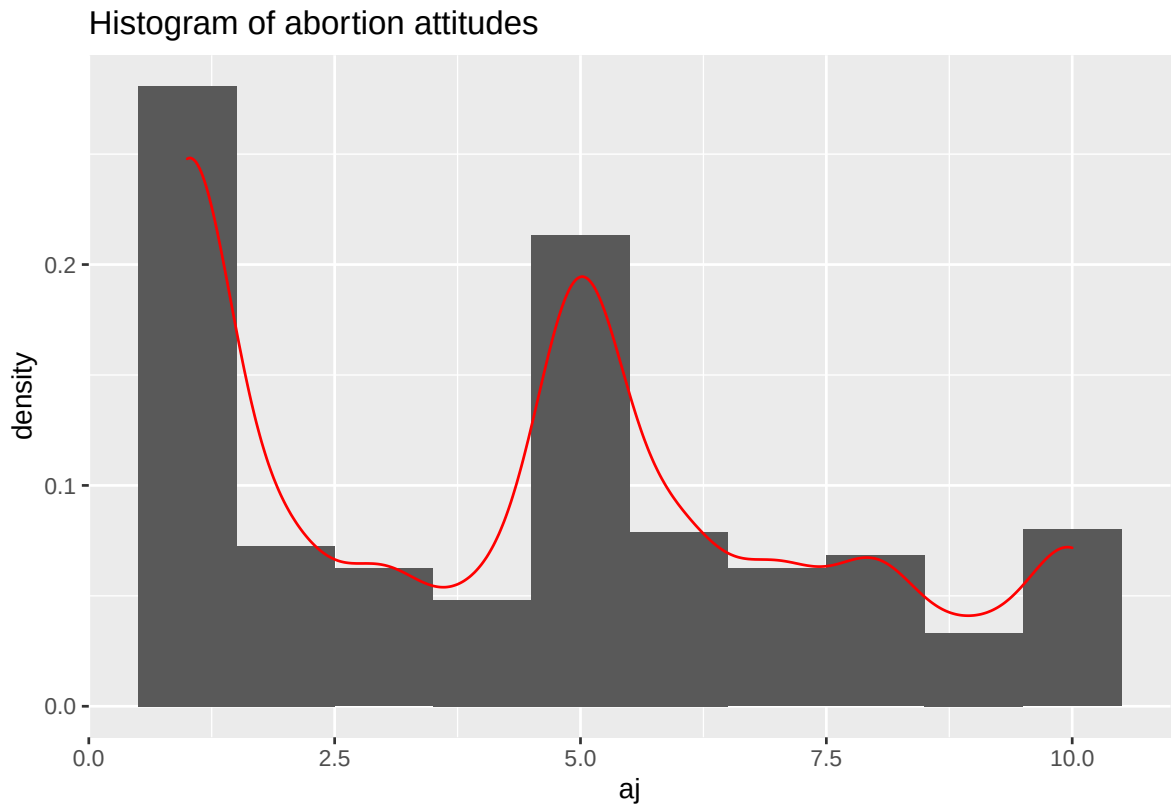
Scrutinizing the data, we see that many survey questions were added only from the 1995 wave of the survey onward. To encompass all the predictor variables of the data set, we decided to remove the observations in generational waves before 1995 so that our data includes all the survey questions.

Moreover, the `wvsccode` for all observations is 840 because all surveys were conducted in the USA. We are hence removing it from our dataset.

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

Before we select our predictor variables, we create a visualization and summary statistics for the response variable.

```r
ggplot(abortion_data, aes(x = aj)) +
  geom_histogram(binwidth = 1, aes(y=..density..)) +
  geom_density(color = "red") +
  labs(title = "Histogram of abortion attitudes")
```



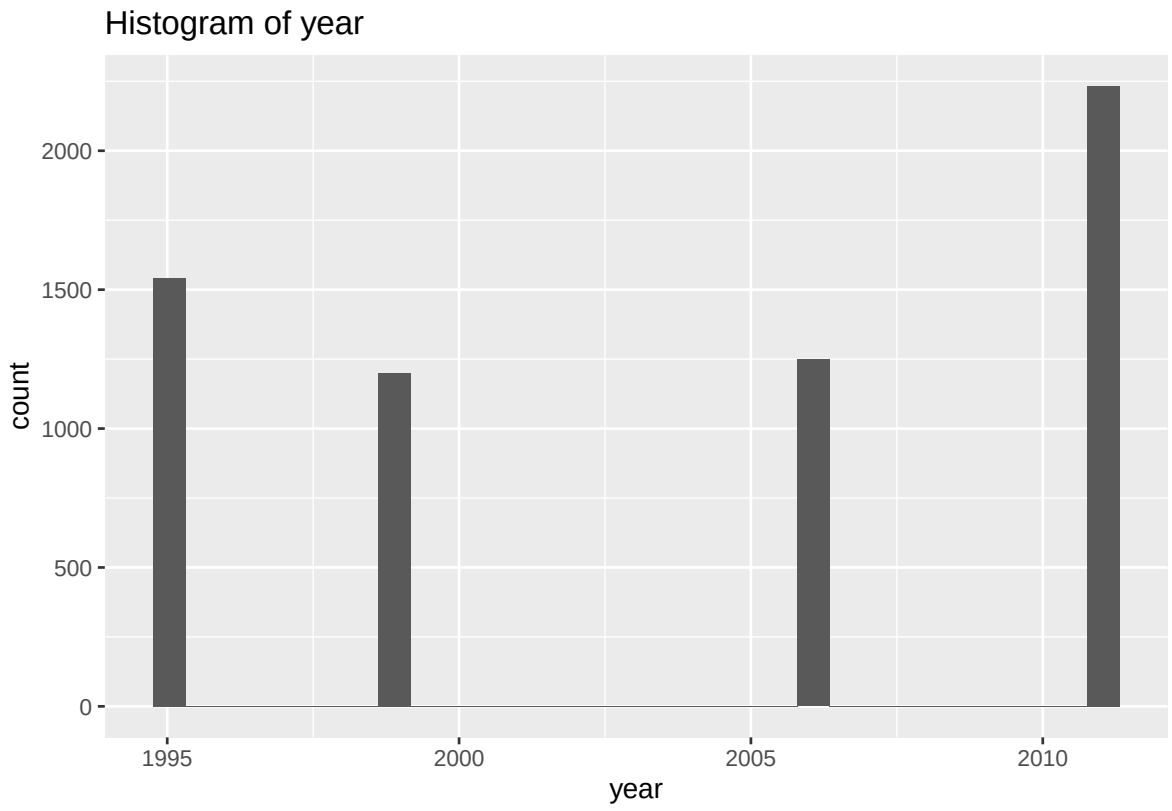

```r
summary(abortion_data$aj)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.000   5.000   4.428   6.000  10.000     214
```

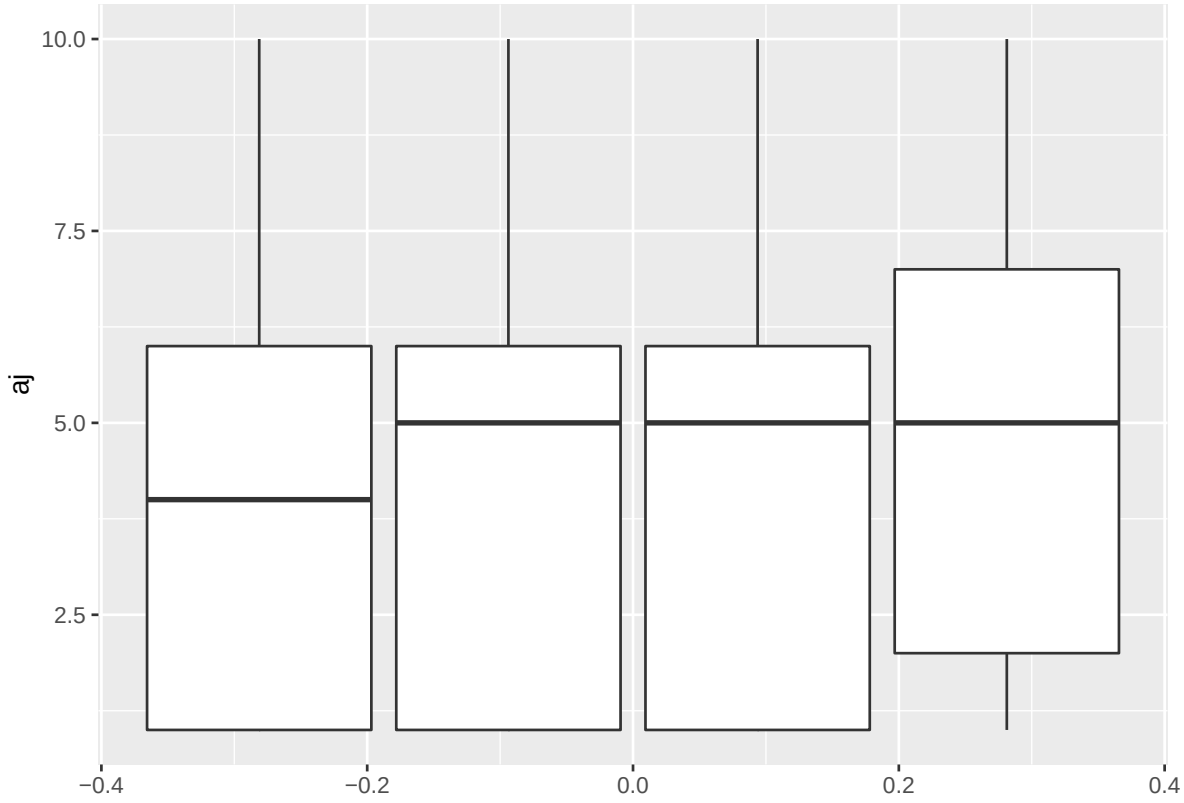
We can observe from this histogram that the distribution of the outcome variable is not a bell shape, and it is trimodal. This is likely because the question's phrasing is similar to a yes/no question, but respondents were asked to give their level of agreement on a scale of 1-10. This may result in our model not being a good fit for the data if we attempt a multiple linear regression model. We have two backup plans for this, if our MLR eventually has a poor performance. First, we can truncate this data into a categorical outcome variable such as (Agree, Disagree, Undecided), and conduct a binomial or multinomial logistic regression. Second, we can filter our population based on various characteristics (if we have good reason to do so) so that our outcome data follows a bell shape.

We now conduct exploratory data analysis (EDA) on each of the 15 predictor variables in the data set. The EDA for each variable comprises a histogram and a simple linear regression (SLR) against the outcome.

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



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

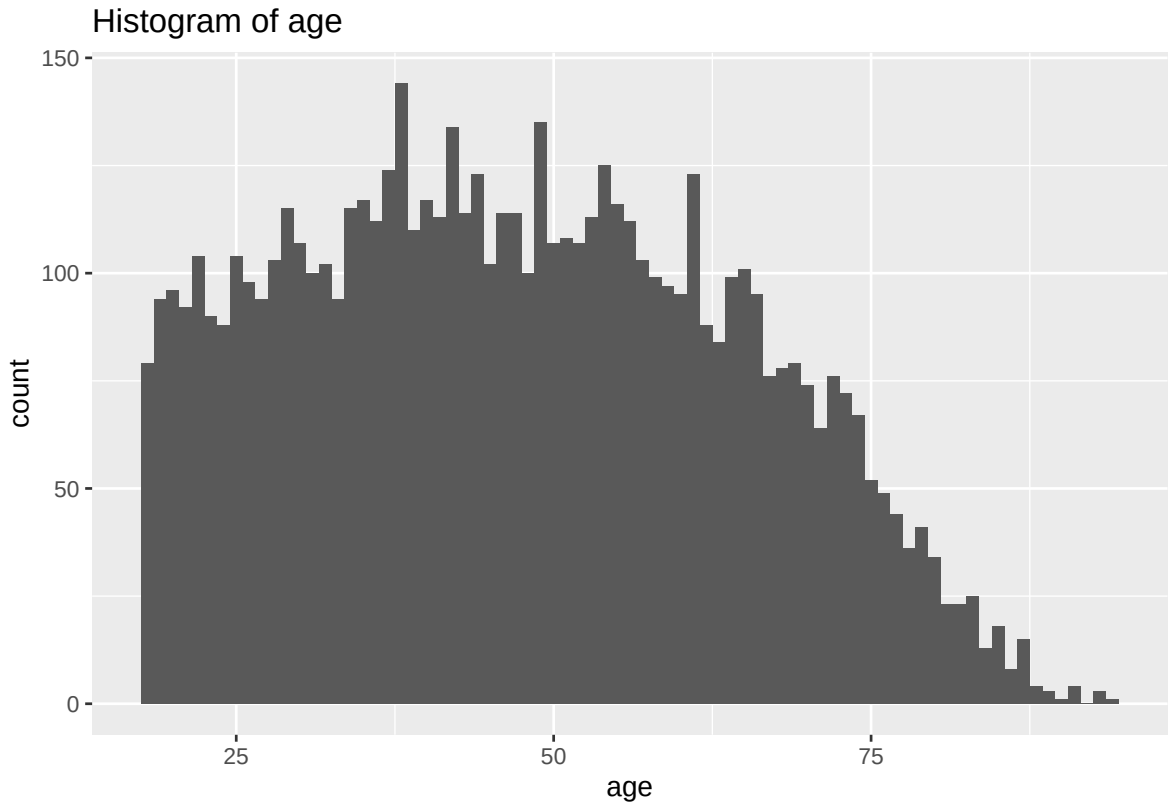



```
#Age
abortion_data <- abortion_data %>%
  mutate(agegroup = case_when(age >= 90 ~ '9',
                                          age >= 80  & age <= 89 ~ '8',
                                          age >= 70  & age <= 79 ~ '7',
                                          age >= 60  & age <= 69 ~ '6',
                                          age >= 50  & age <= 59 ~ '5',
                                          age >= 40  & age <= 49 ~ '4',
                                          age >= 30  & age <= 39 ~ '3',
                                          age >= 20  & age <= 29 ~ '2',
                              age <= 19 ~ '1'))
abortion_data
```

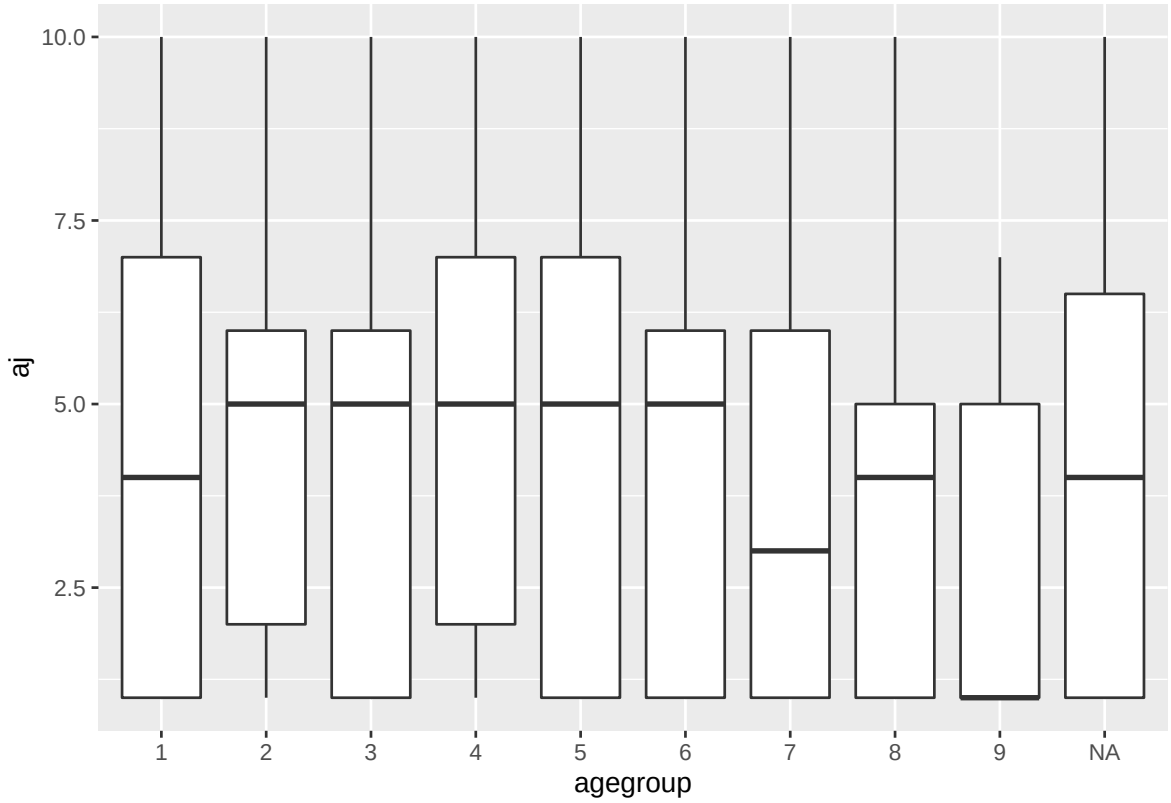
```
# A tibble: 6,223 x 15
```

```
   year    aj   age collegeed female unemployed ideology satisfinancial postma4
  <dbl> <dbl> <dbl>     <dbl>  <dbl>      <dbl>    <dbl>          <dbl>   <dbl>
 1  1995     1    44         0      0          0        2              9      -1
 2  1995     7    40         0      1          0        4              8       0
 3  1995    10    43         0      0          0        3              2       1
 4  1995     1    36         0      1          0        7              9       0
 5  1995    10    25         1      1          0        5              2       1
 6  1995     7    39         1      0          0        4              9       0
 7  1995     1    80         0      1          0       NA              4       0
 8  1995     1    48         0      1          0       10              4       1
 9  1995     1    93         0      1          0        8             10       0
10  1995     1    32         0      0          0        7              9       0
# ... with 6,213 more rows, and 6 more variables: cai <dbl>,
#   trustmostpeople <dbl>, godimportant <dbl>, respectauthority <dbl>,
#   nationalpride <dbl>, agegroup <chr>
```

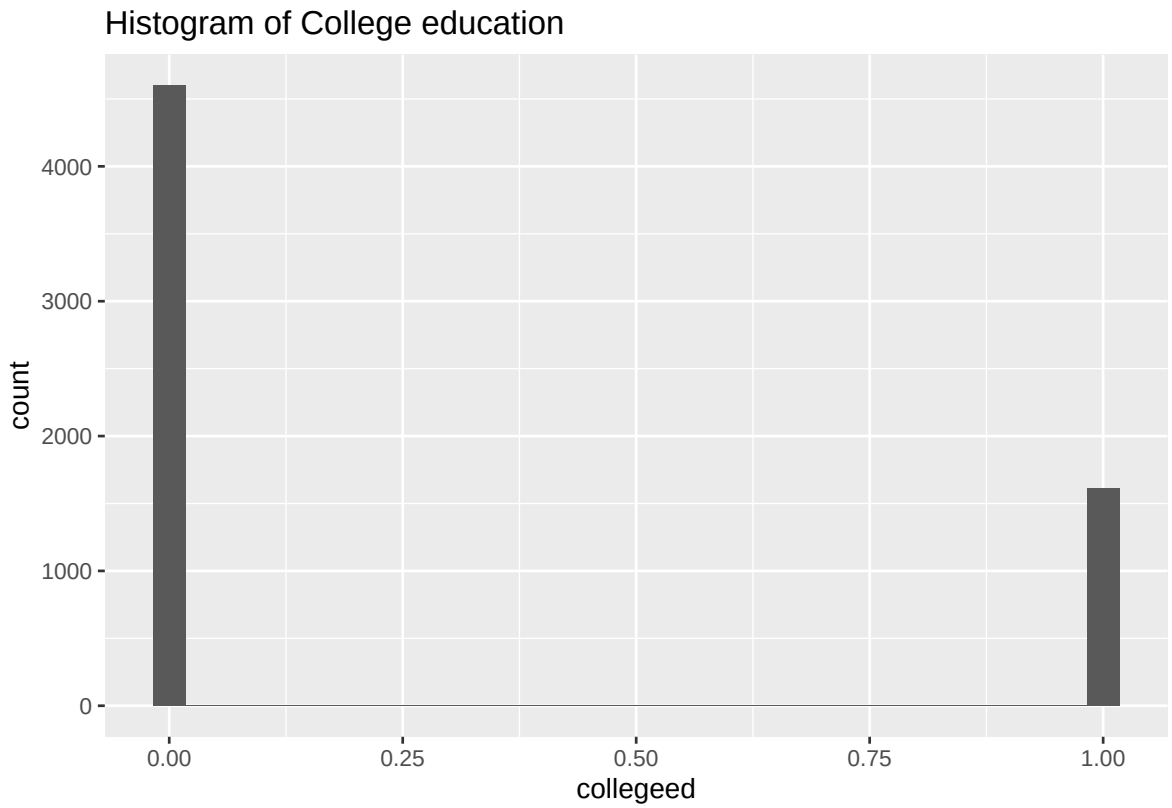
```
ggplot(abortion_data, aes(x = age)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Histogram of age")
```



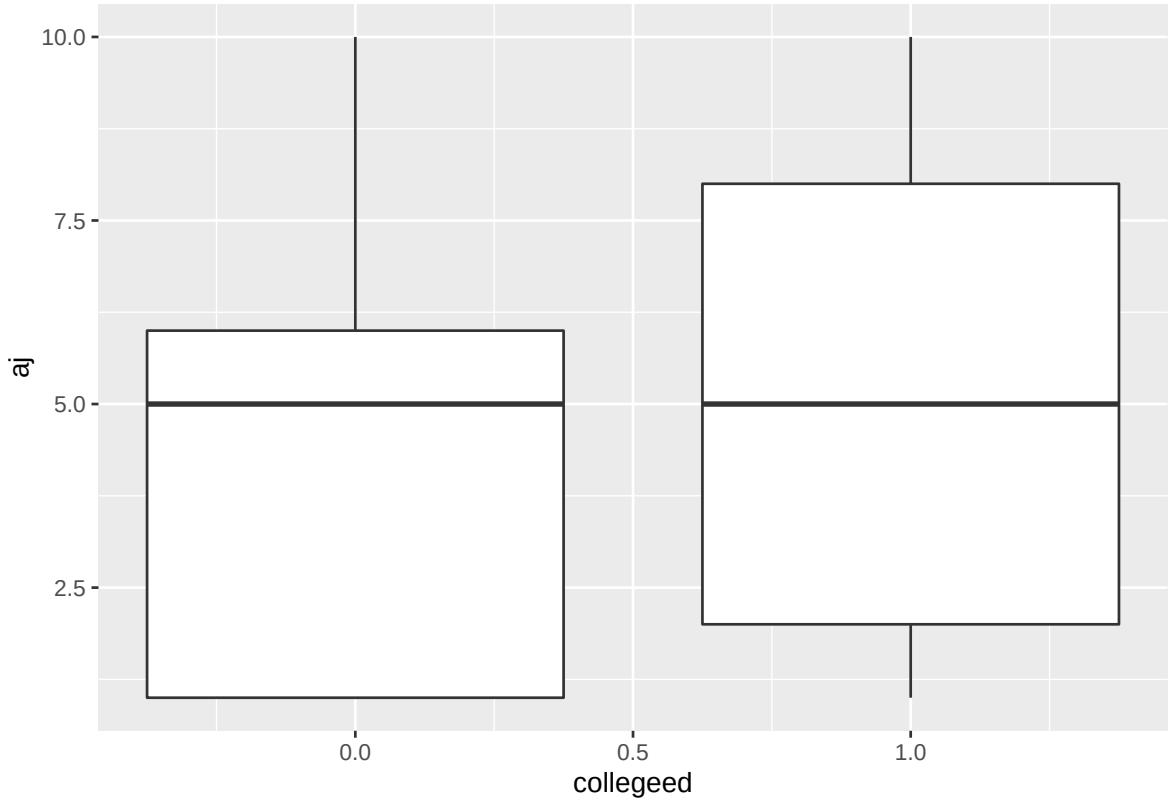

```
ggplot(abortion_data, aes(x = agegroup, group = agegroup, y = aj)) +
  geom_boxplot()
```



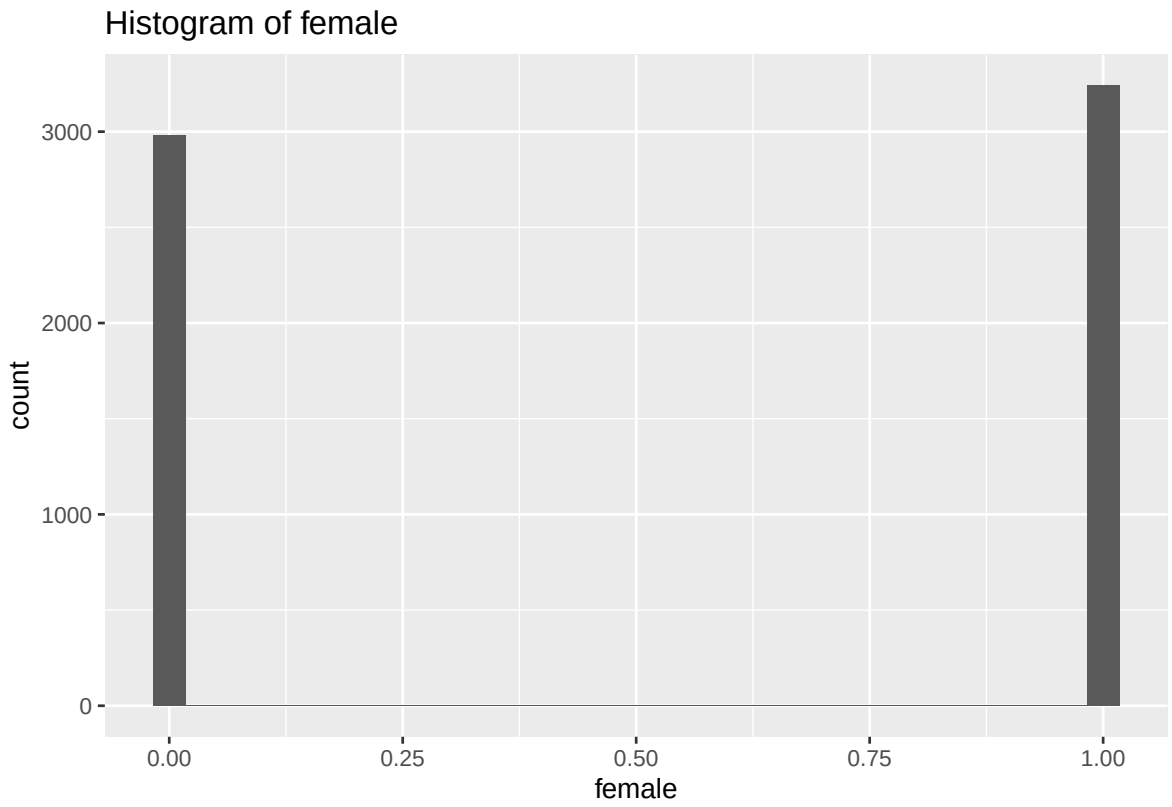

```
# College education
ggplot(abortion_data, aes(x = collegeed)) +
  geom_histogram() +
  labs(title = "Histogram of College education")
```



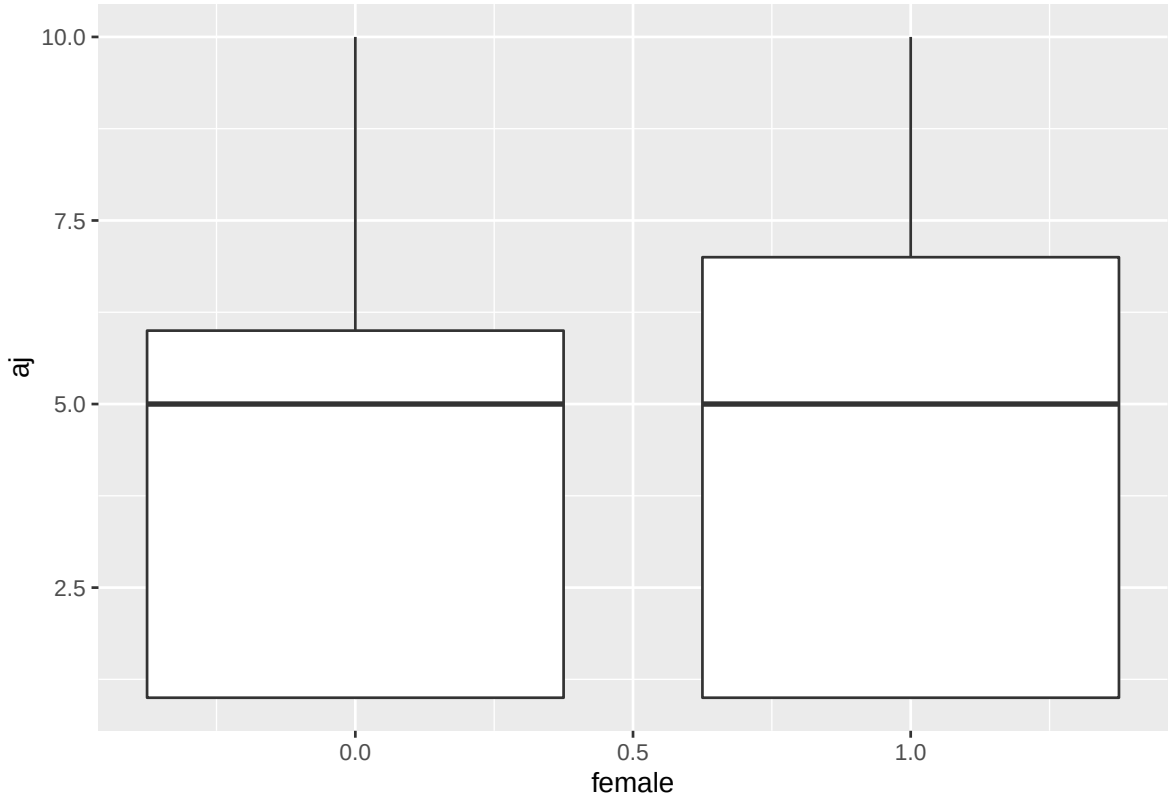

```
ggplot(abortion_data, aes(x = collegeed, group = collegeed, y = aj)) +
  geom_boxplot()
```



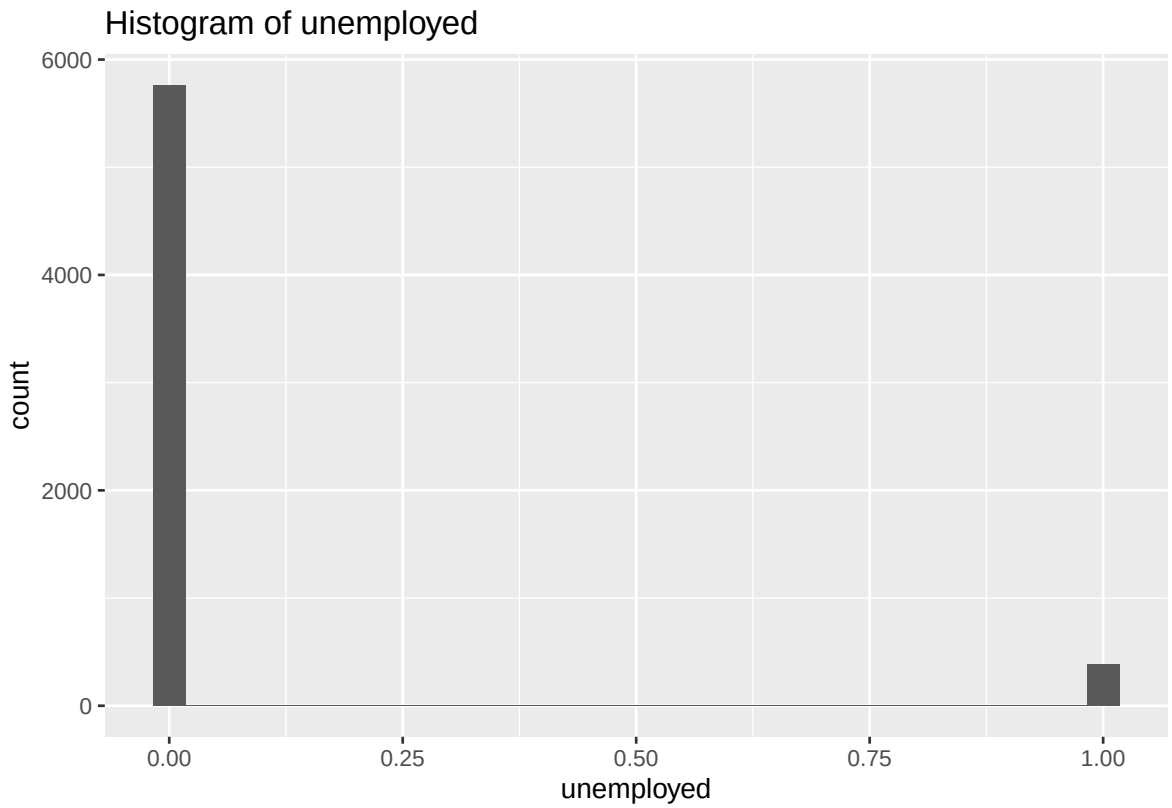

```
# Female
ggplot(abortion_data, aes(x = female)) +
  geom_histogram() +
  labs(title = "Histogram of female")
```



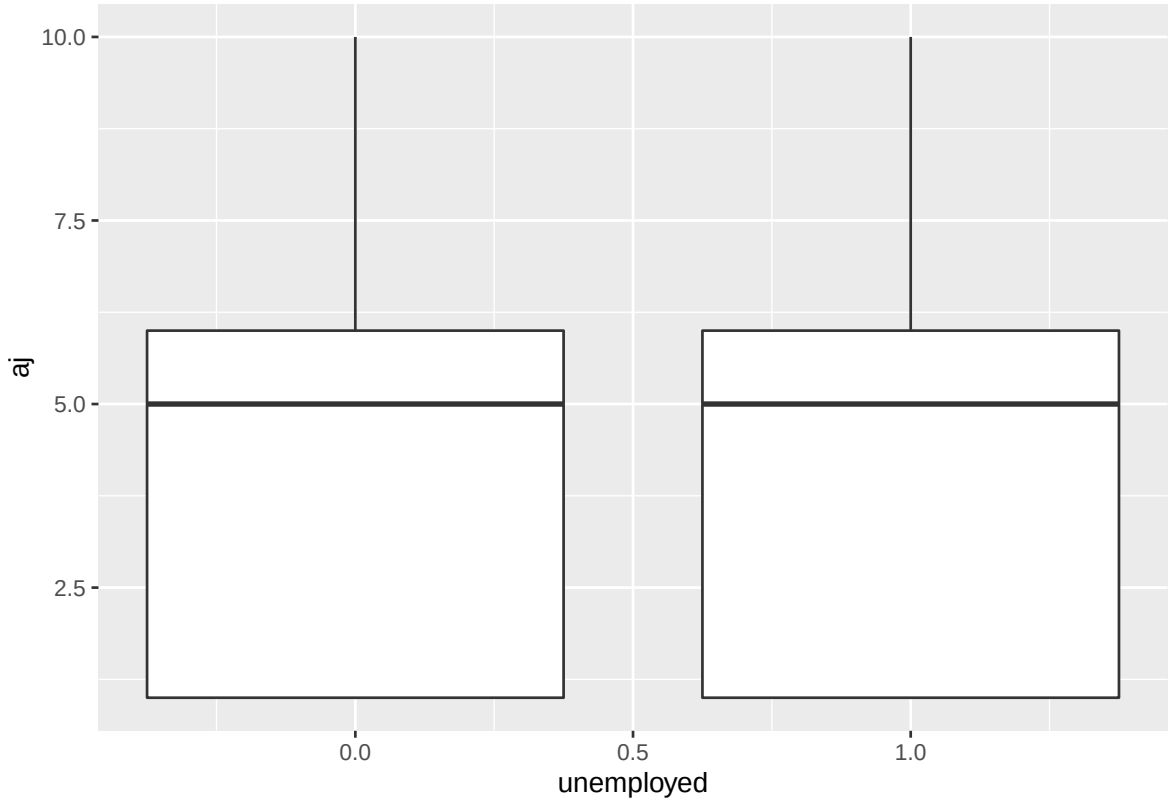

```
ggplot(abortion_data, aes(x = female,group = female, y = aj)) +
  geom_boxplot()
```



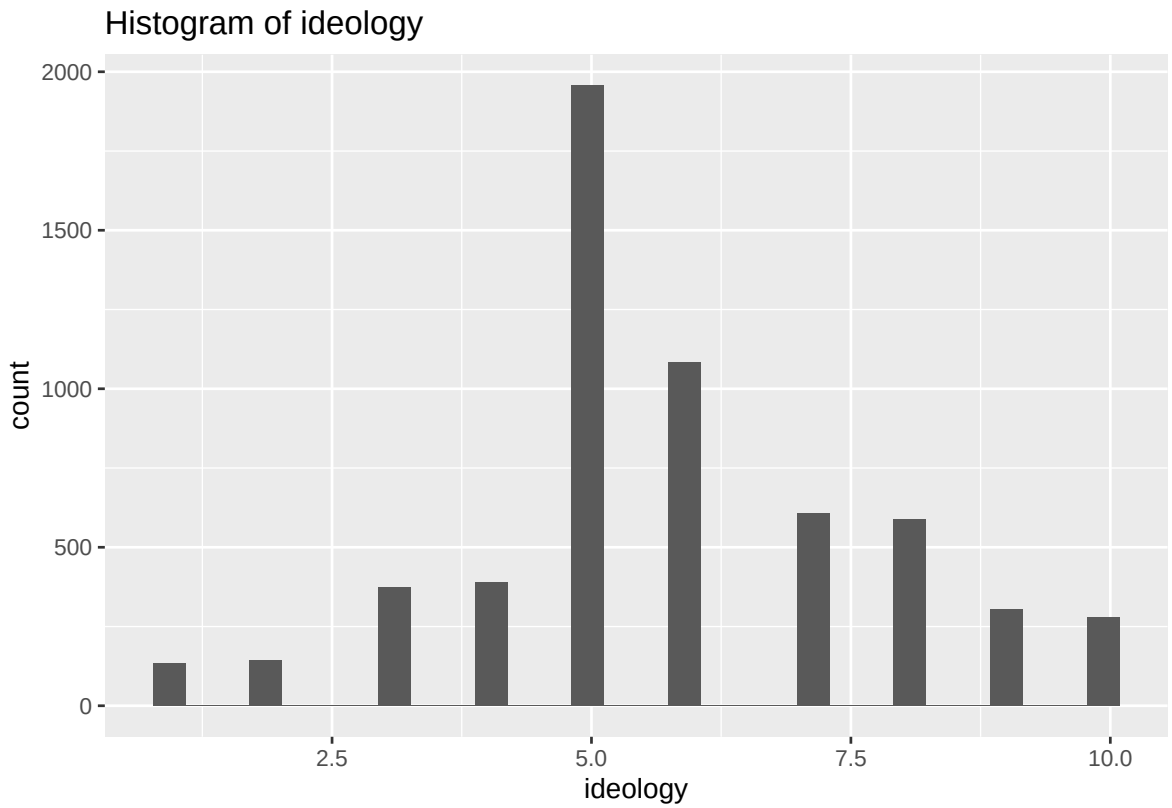

```
# Unemployed
ggplot(abortion_data, aes(x = unemployed)) +
  geom_histogram() +
  labs(title = "Histogram of unemployed")
```



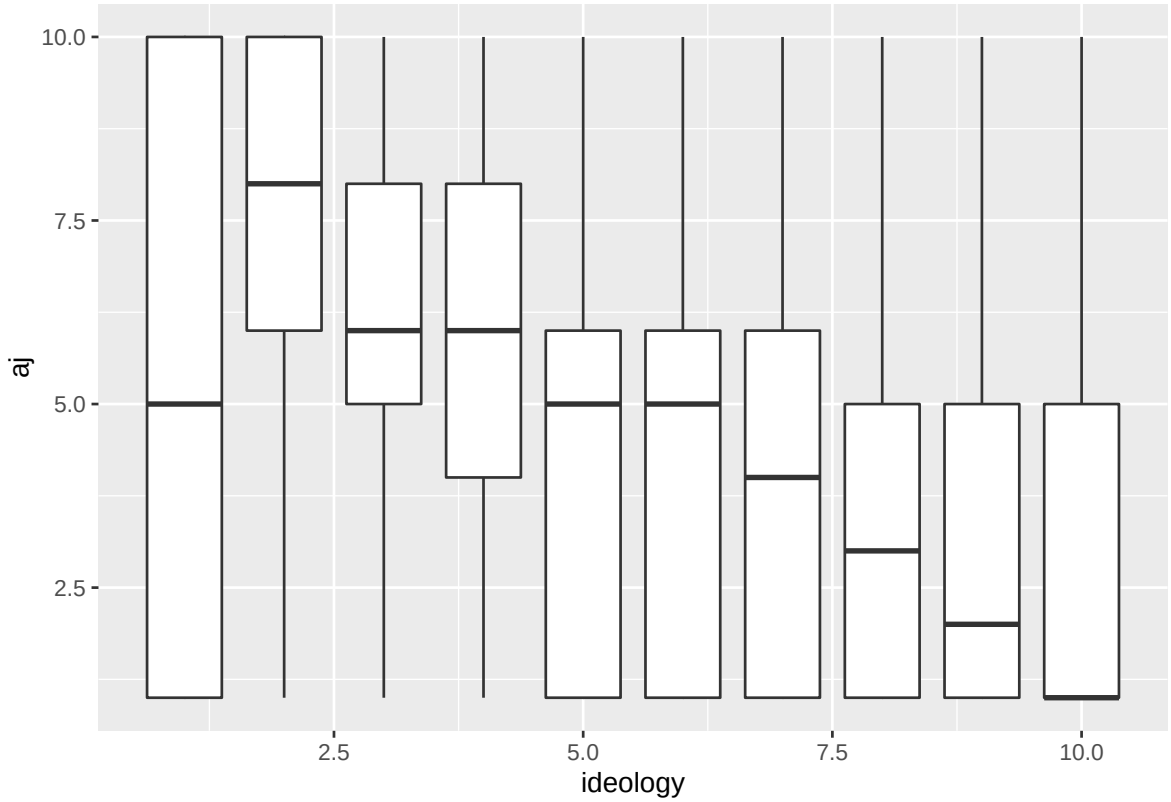

```
ggplot(abortion_data, aes(x = unemployed, group = unemployed, y = aj)) +
  geom_boxplot()
```



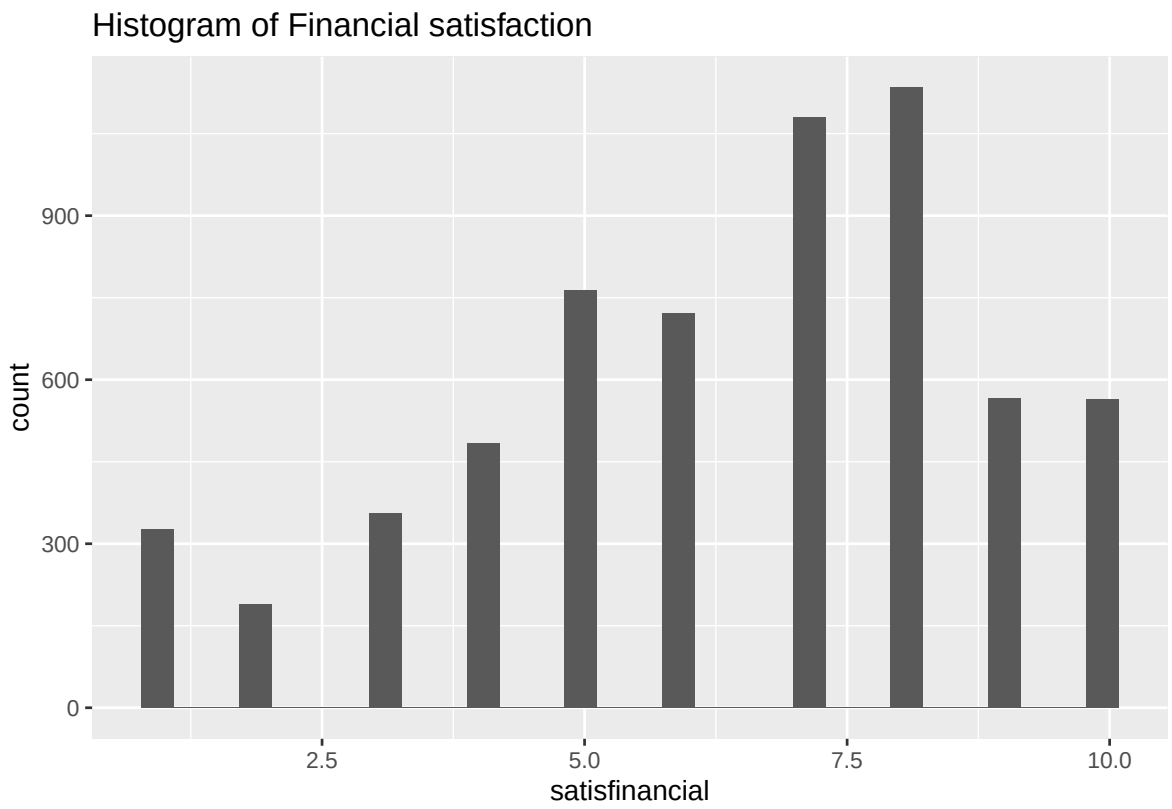

```
# ideology
ggplot(abortion_data, aes(x = ideology)) +
  geom_histogram() +
  labs(title = "Histogram of ideology")
```



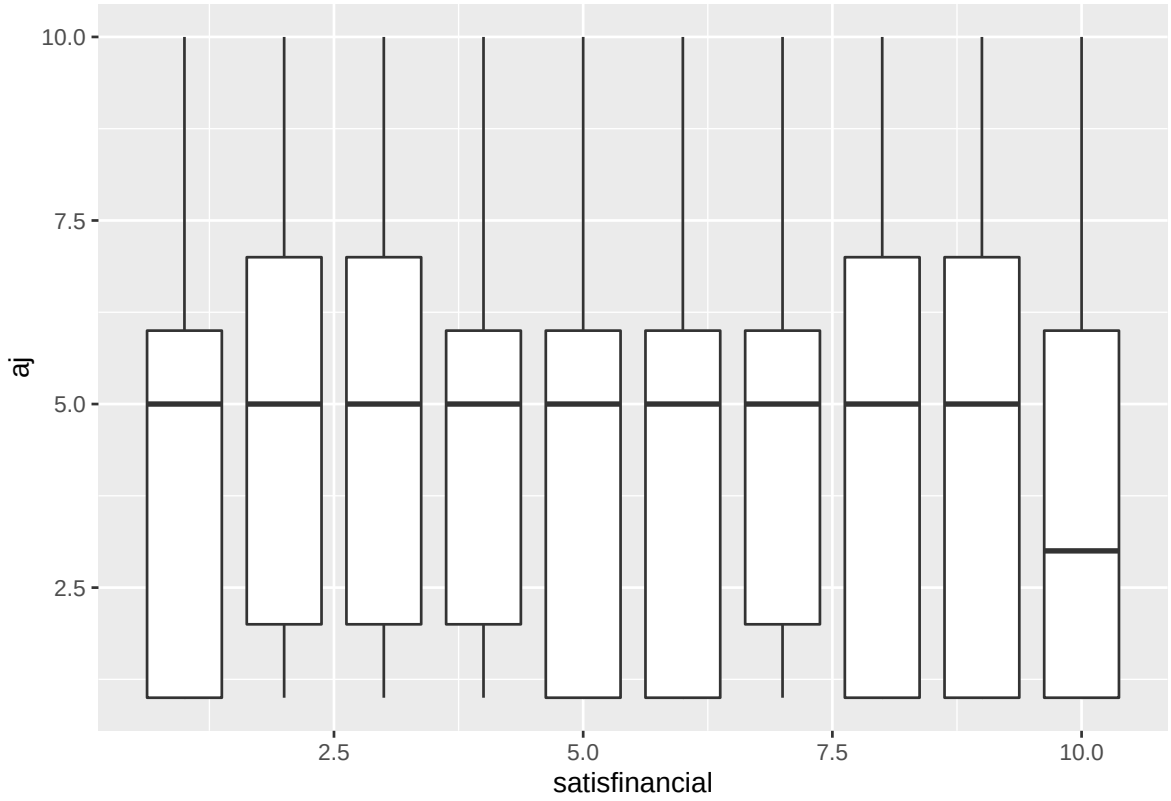

```
ggplot(abortion_data, aes(x = ideology,group = ideology, y = aj)) +
  geom_boxplot()
```



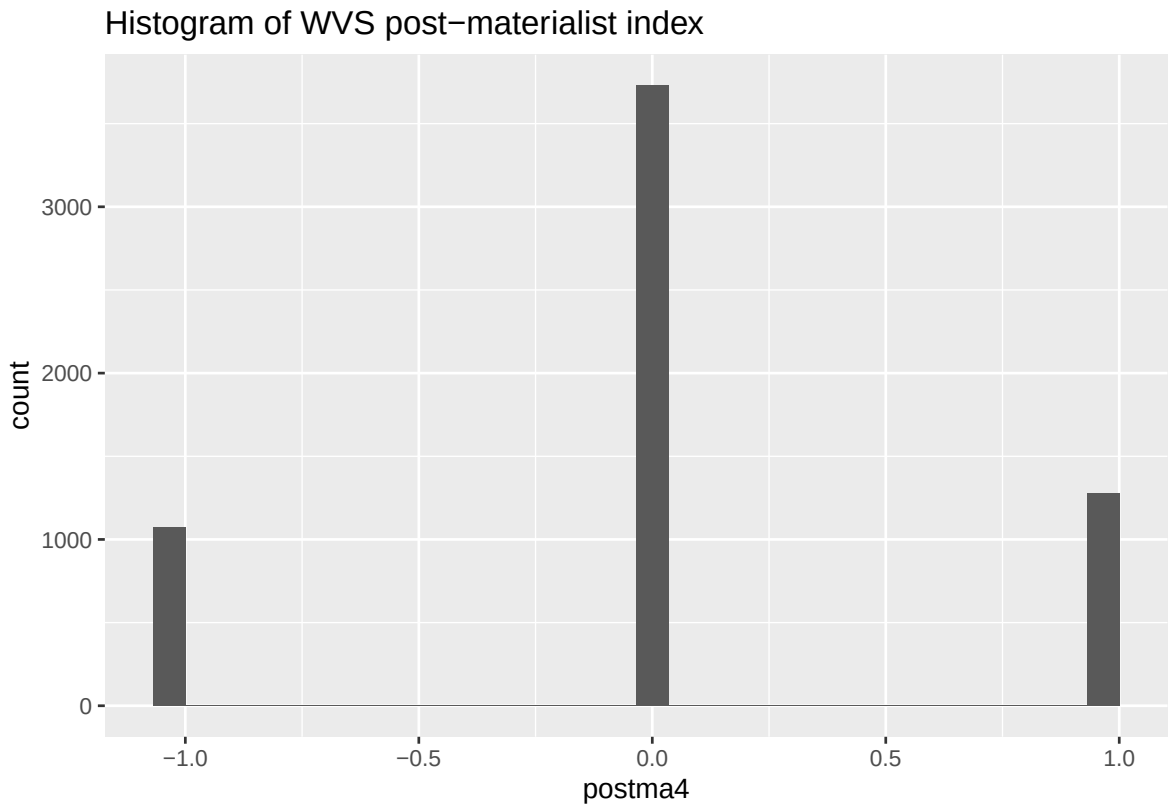

```
# Financial satisfaction
ggplot(abortion_data, aes(x = satisfinancial)) +
  geom_histogram() +
  labs(title = "Histogram of Financial satisfaction")
```



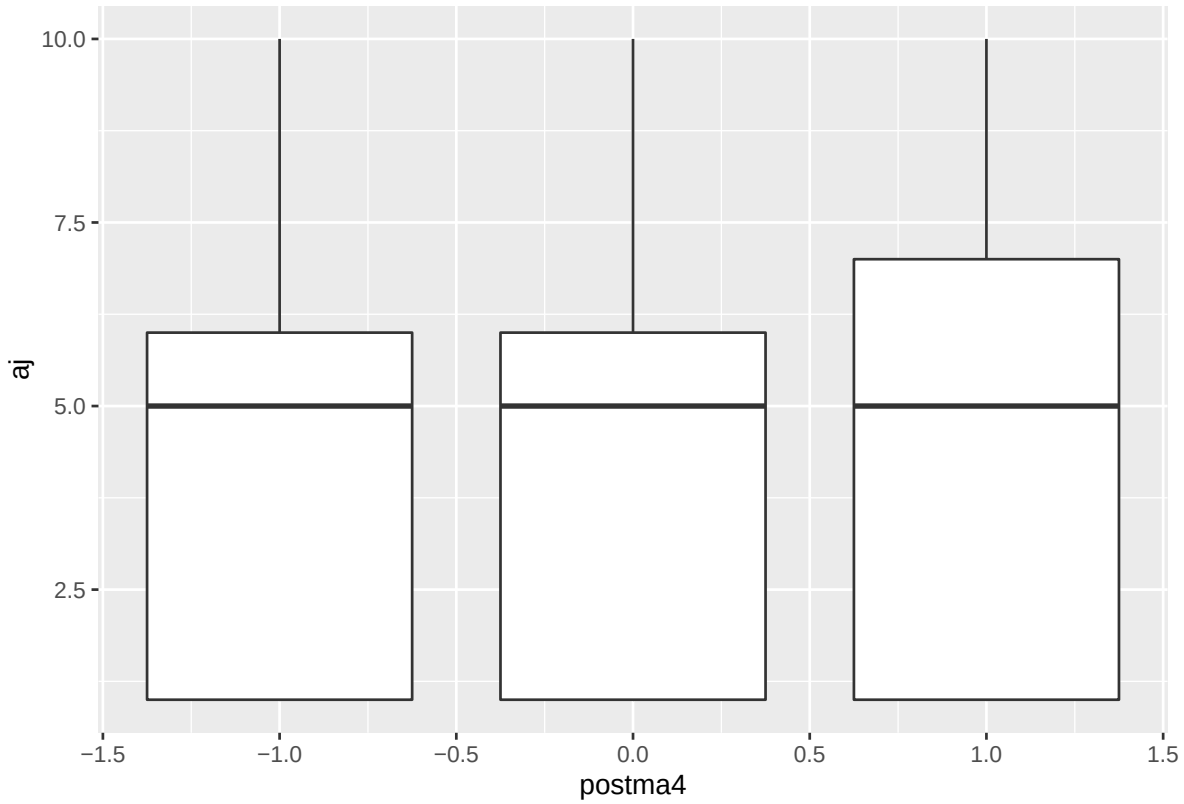

```
ggplot(abortion_data, aes(x = satisfinancial,group = satisfinancial, y = aj)) +
  geom_boxplot()
```



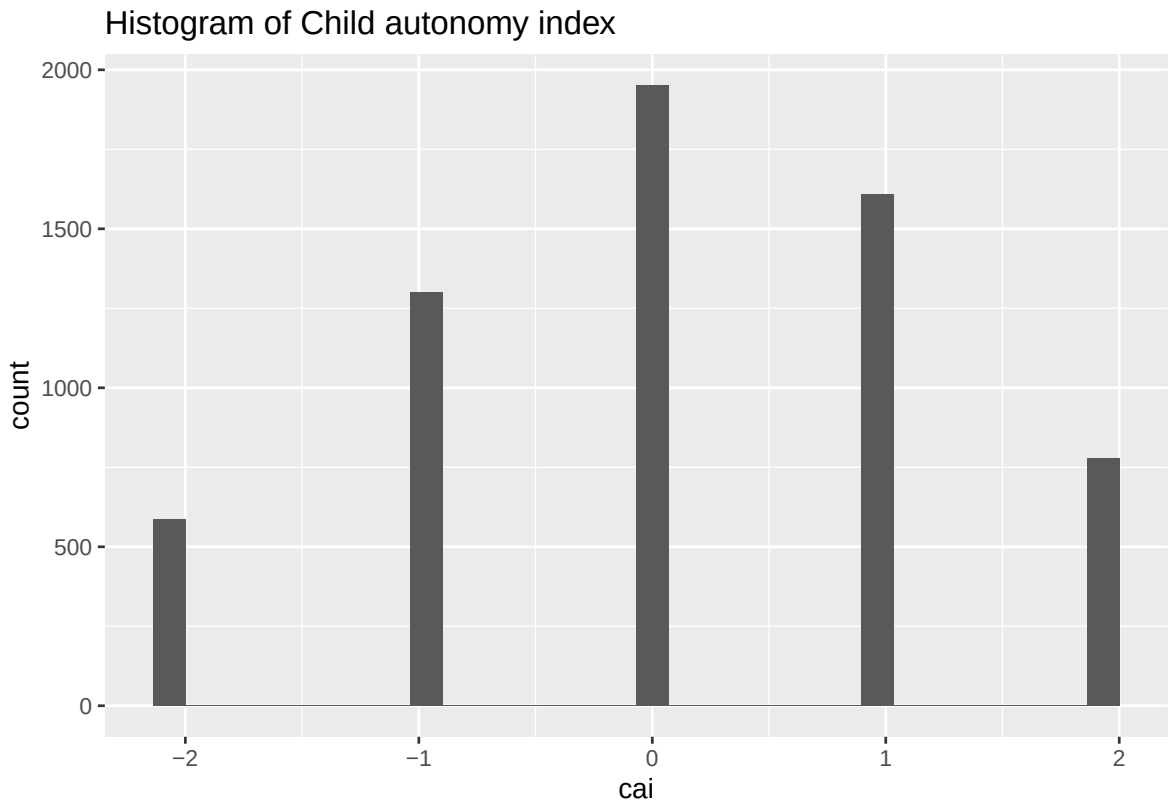

```
# WVS post-materialist index
ggplot(abortion_data, aes(x = postma4)) +
  geom_histogram() +
  labs(title = "Histogram of WVS post-materialist index")
```



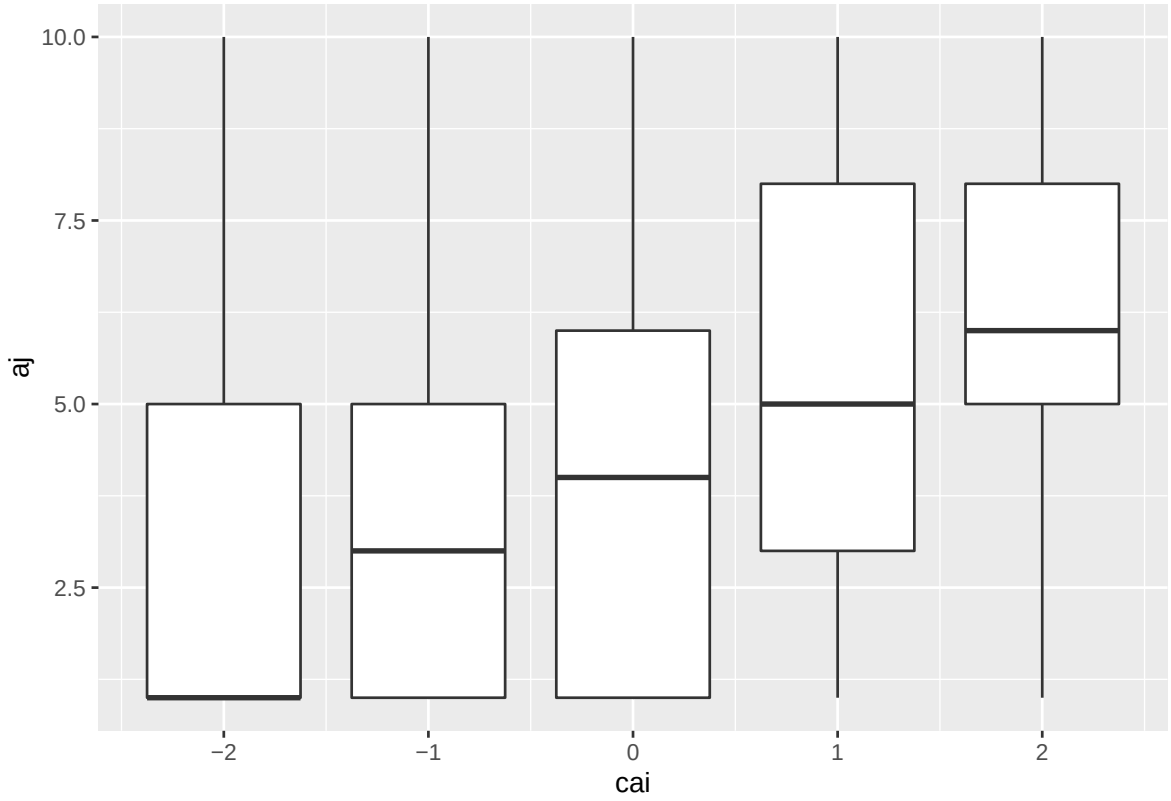

```
ggplot(abortion_data, aes(x = postma4, group = postma4,y = aj)) +
  geom_boxplot()
```



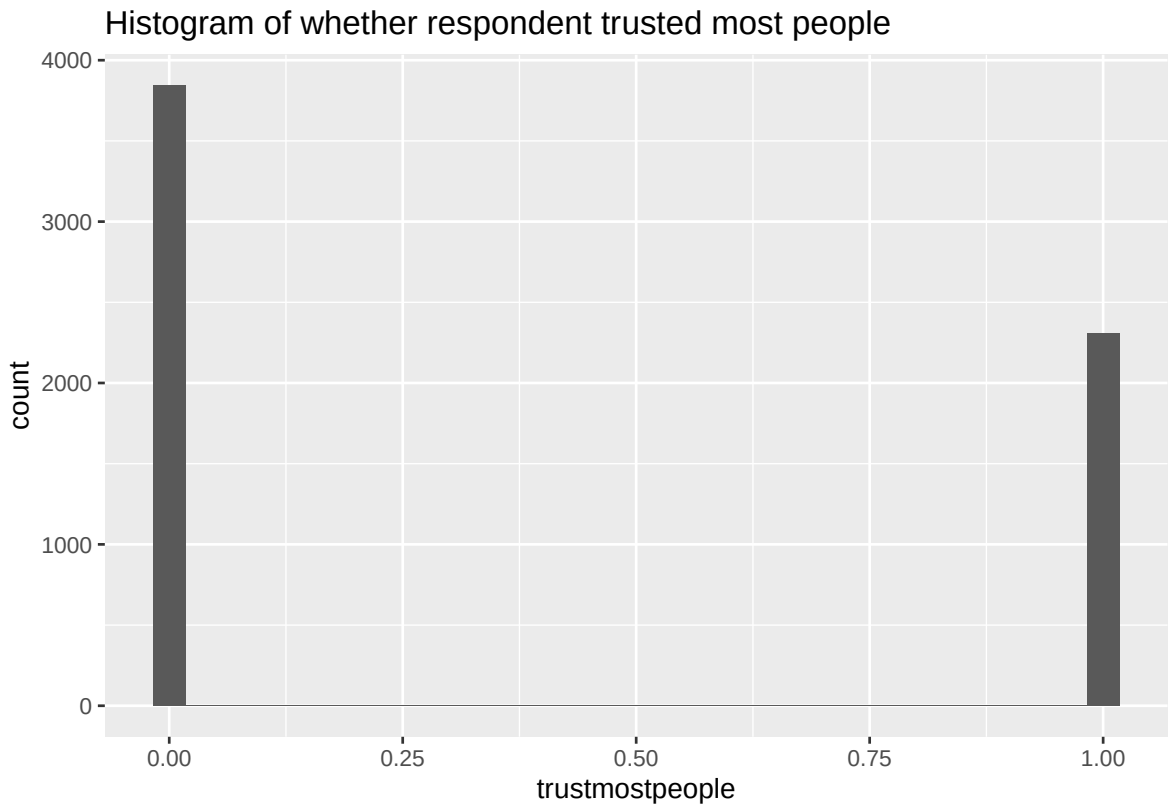

```
# Child autonomy index
ggplot(abortion_data, aes(x = cai)) +
  geom_histogram() +
  labs(title = "Histogram of Child autonomy index")
```



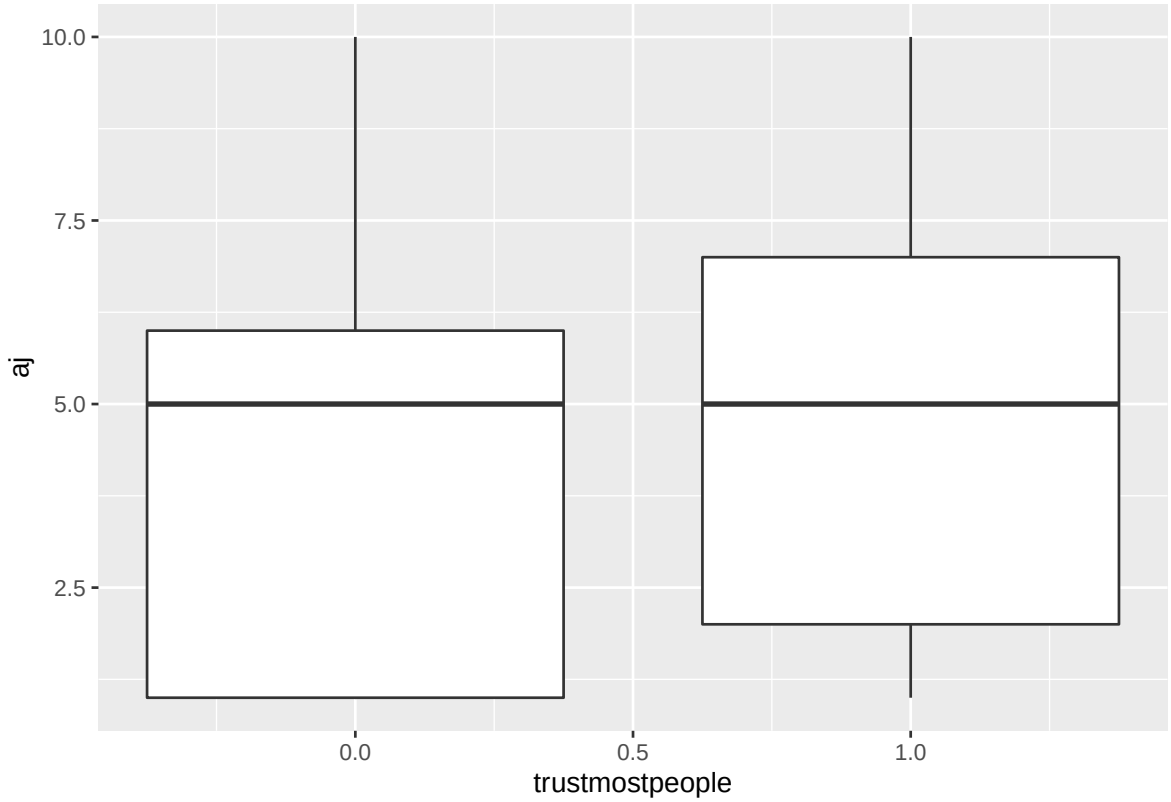

```
ggplot(abortion_data, aes(x = cai, group = cai, y = aj)) +
  geom_boxplot()
```



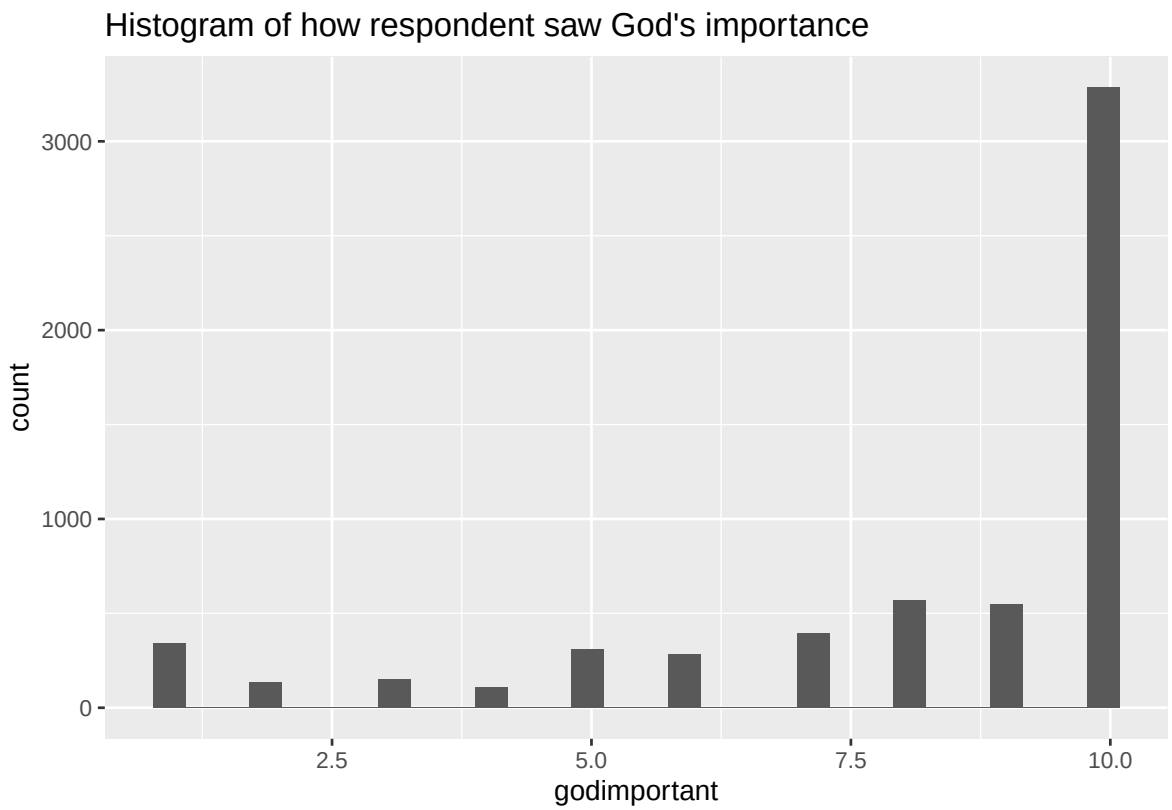

```
# Trust most people
ggplot(abortion_data, aes(x = trustmostpeople)) +
  geom_histogram() +
  labs(title = "Histogram of whether respondent trusted most people")
```



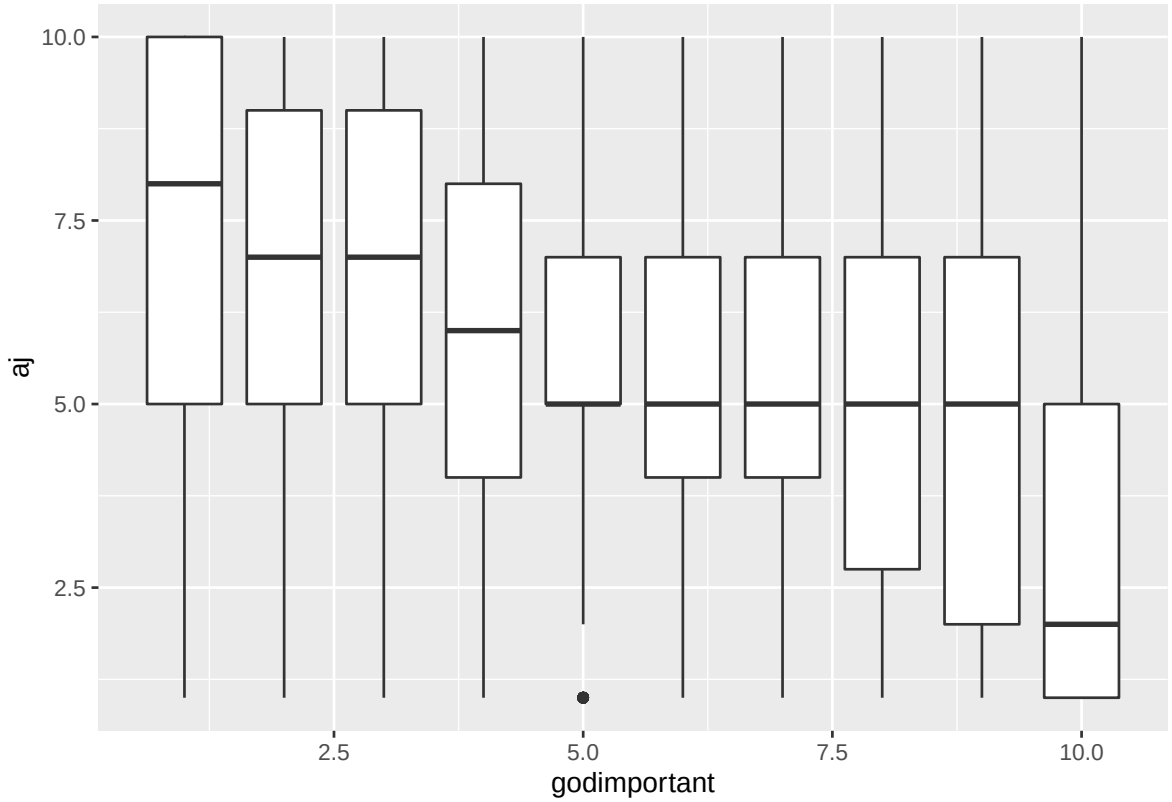

```
ggplot(abortion_data, aes(x = trustmostpeople, group = trustmostpeople, y = aj)) +
  geom_boxplot()
```



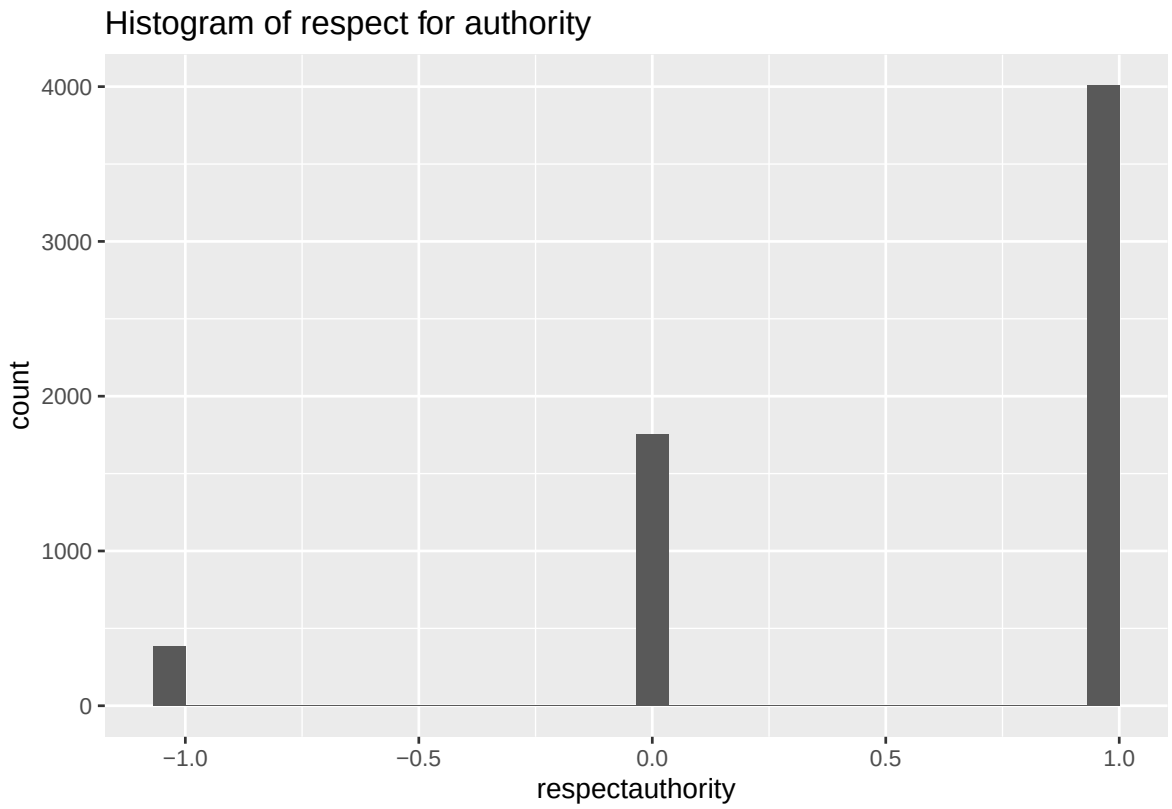

```
# Importance of God
ggplot(abortion_data, aes(x = godimportant)) +
  geom_histogram() +
  labs(title = "Histogram of how respondent saw God's importance")
```



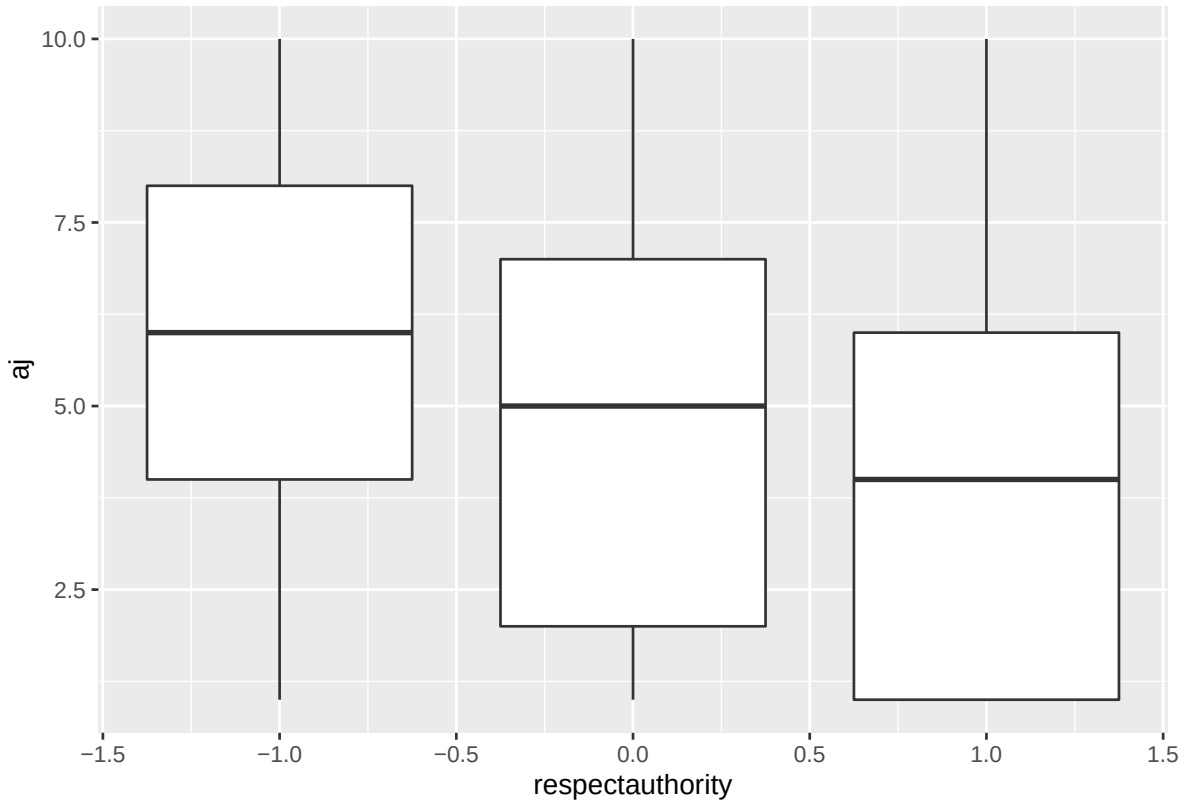

```
ggplot(abortion_data, aes(x = godimportant, group = godimportant, y = aj)) +
  geom_boxplot()
```



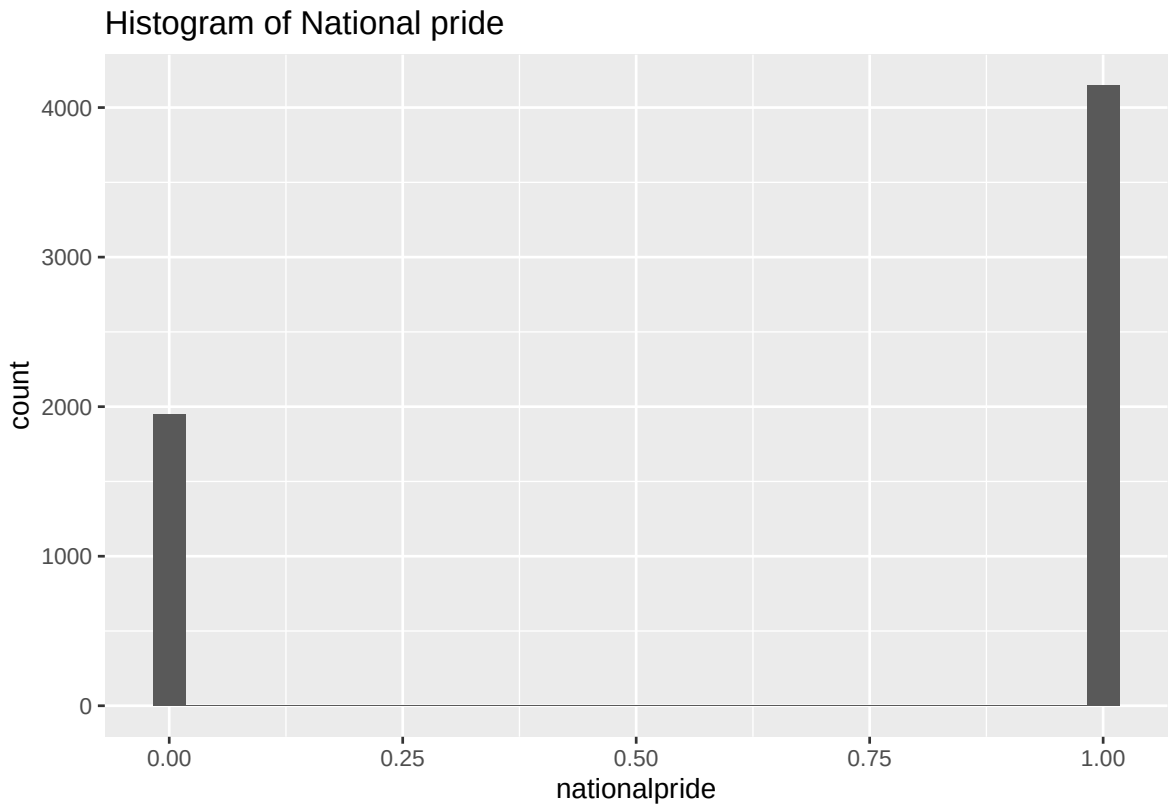

```
# Respect for authority
ggplot(abortion_data, aes(x = respectauthority)) +
  geom_histogram() +
  labs(title = "Histogram of respect for authority")
```



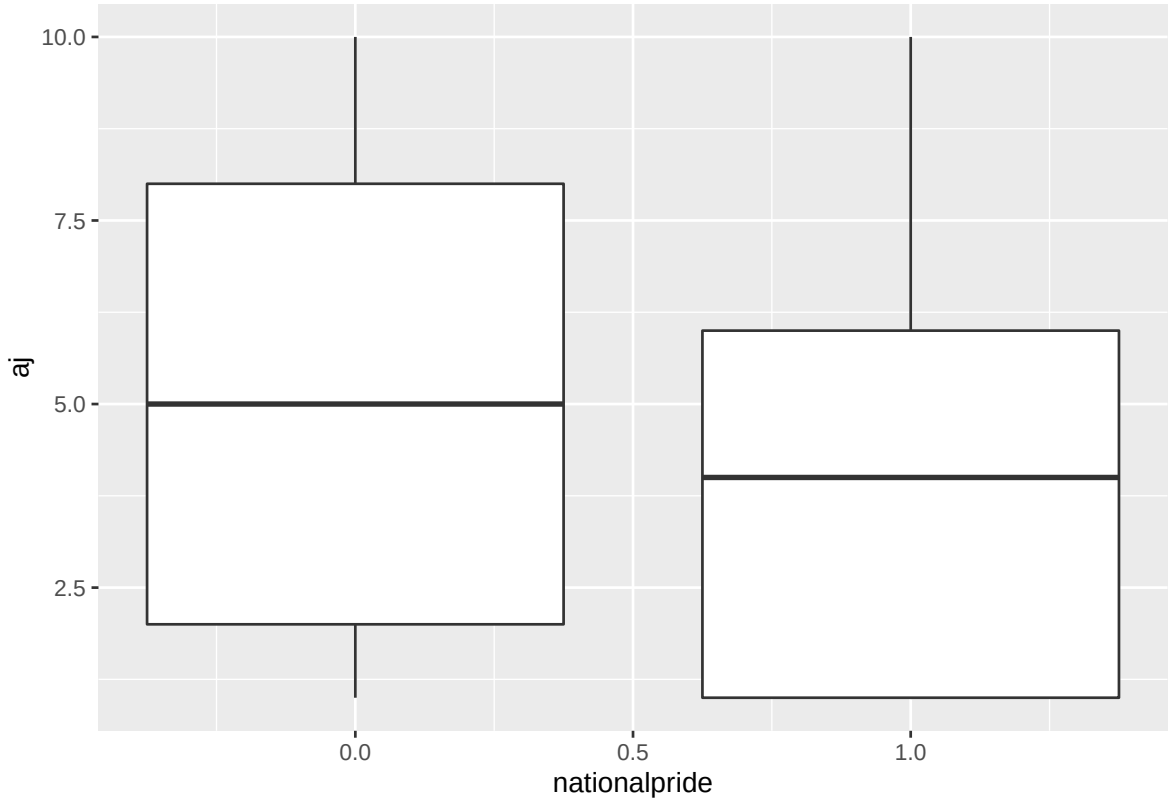

```
ggplot(abortion_data, aes(x = respectauthority, group = respectauthority, y = aj)) +
  geom_boxplot()
```




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

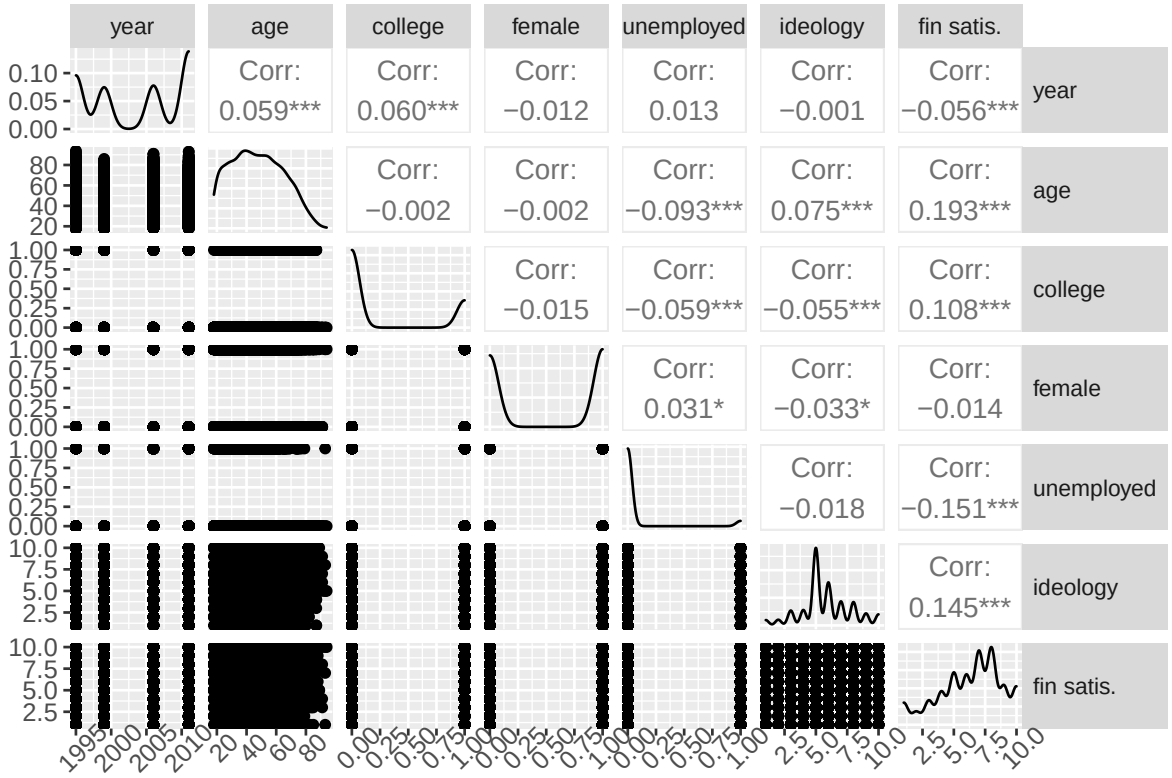



```
ggplot(abortion_data, aes(x = nationalpride, group = nationalpride, y = aj)) +
  geom_boxplot()
```

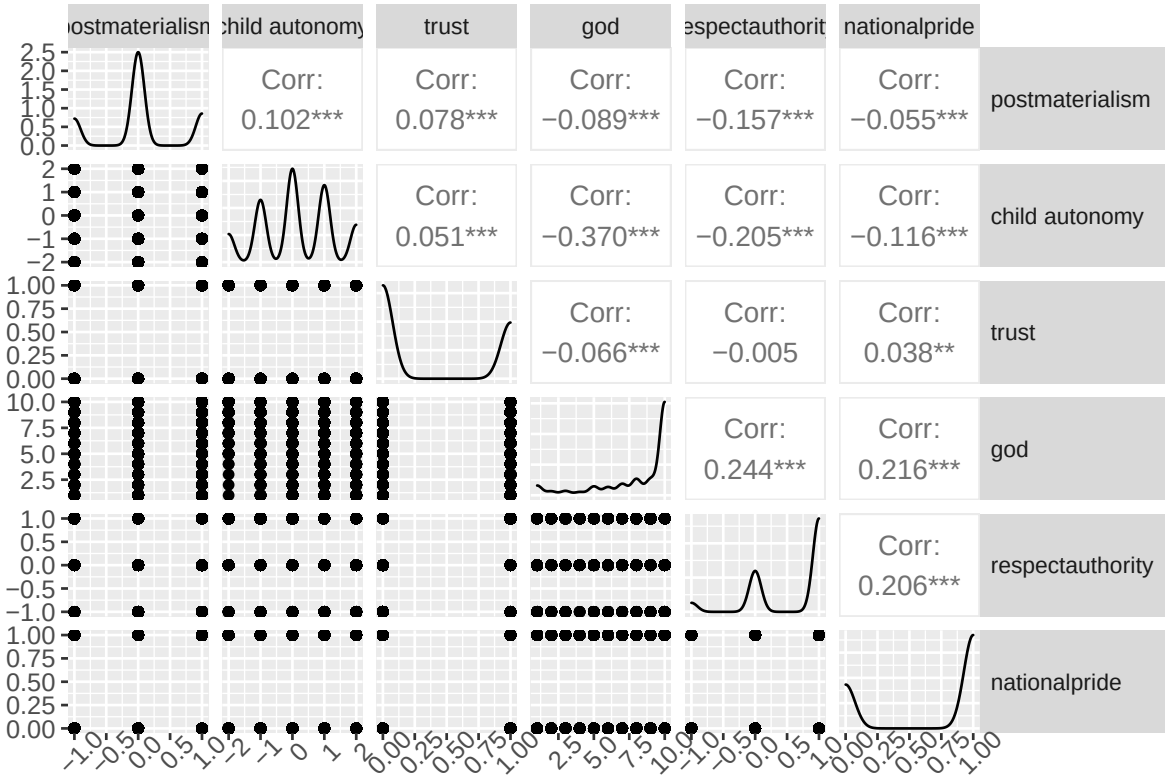



We now test if any of the predictors are strongly correlated with each other.

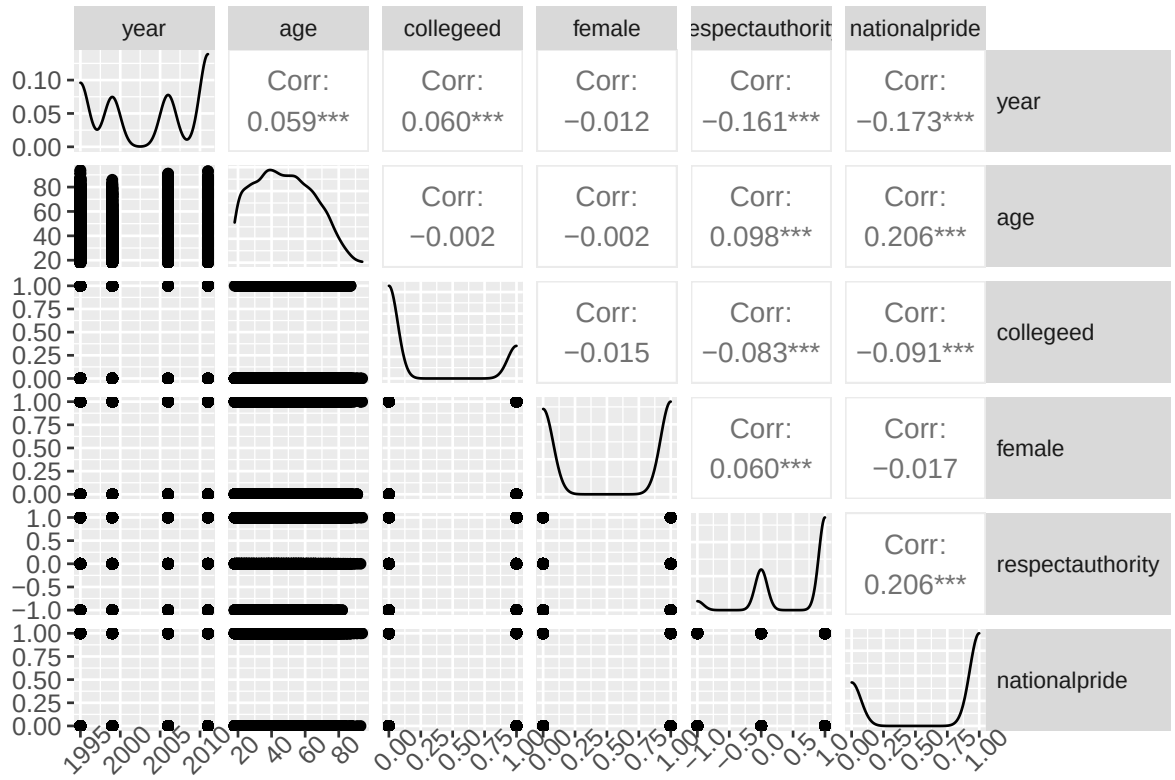
```
#I am doing ggpairs on 13 predictor variables.
#To ensure they fit into the screen, I will make 4 matrices with subsets

ggpairs(abortion_data,
  columns = c("year", "age", "collegeed", "female", "unemployed",
              "ideology", "satisfinancial"),
  columnLabels = c("year", "age", "college", "female",
                   "unemployed", "ideology", "fin satis.")) +
   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)
    )
```



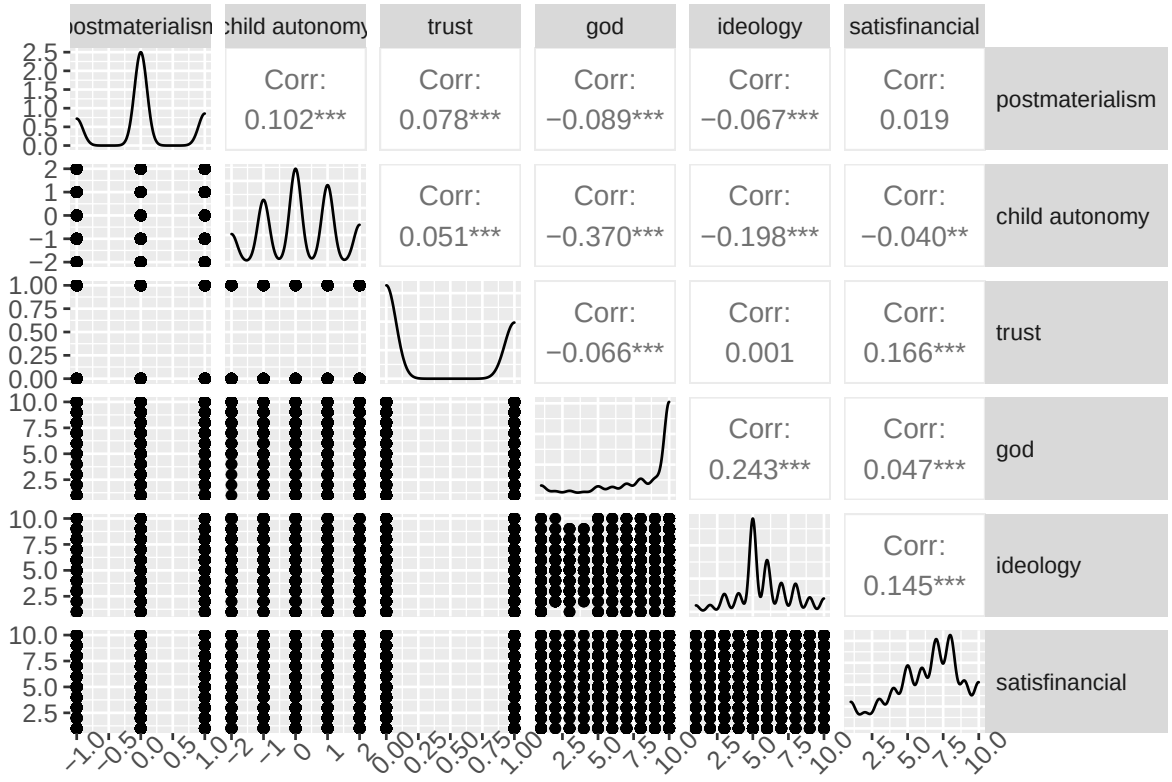

```
ggpairs(abortion_data,
  columns = c("postma4", "cai", "trustmostpeople", "godimportant",
              "respectauthority","nationalpride"),
  columnLabels = c("postmaterialism", "child autonomy", "trust", "god",
              "respectauthority","nationalpride")) +
   theme(
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(angle = 45, size = 10),
    strip.text.y = element_text(angle = 0, hjust = 0)
    )
```




```
ggpairs(abortion_data,
  columns = c("year", "age", "collegeed", "female",
              "respectauthority","nationalpride"),
  columnLabels = c("year", "age", "collegeed", "female",
              "respectauthority","nationalpride")) +
   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)
    )
```




```
ggpairs(abortion_data,
  columns = c("postma4", "cai", "trustmostpeople", "godimportant",
              "ideology", "satisfinancial"),
  columnLabels = c("postmaterialism", "child autonomy", "trust", "god",
              "ideology", "satisfinancial")) +
   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)
    )
```




From these correlation matrices, we can conclude that the highest correlations (above 0.2) are those between `godimportant` and `cai` 0.370 `godimportant` and `respectauthority` 0.244 `godimportant` and `ideology` 0.243 `godimportant` and `nationalpride` 0.216 `nationalpride` and `age` 0.206 `nationalpride` and `respectauthority` 0.206 `respectauthority` and `cai` 0.205

We want to choose predictor variables that are not strongly correlated with each other, are relatively balanced in our data set, and visually appear to have a significant relationship with the outcome. Now that we have calculated the correlation matrix and know that some variables may have multicollinearity, we will select variables that do not exhibit significant multicollinearity for our model. Moreover, we can run several multiple linear regression models with different combinations of predictors to see which model performs best and is least affected by multicollinearity.

## Data dictionary

The data dictionary can be found here.