Draft

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

Paste-updated-content-here

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.

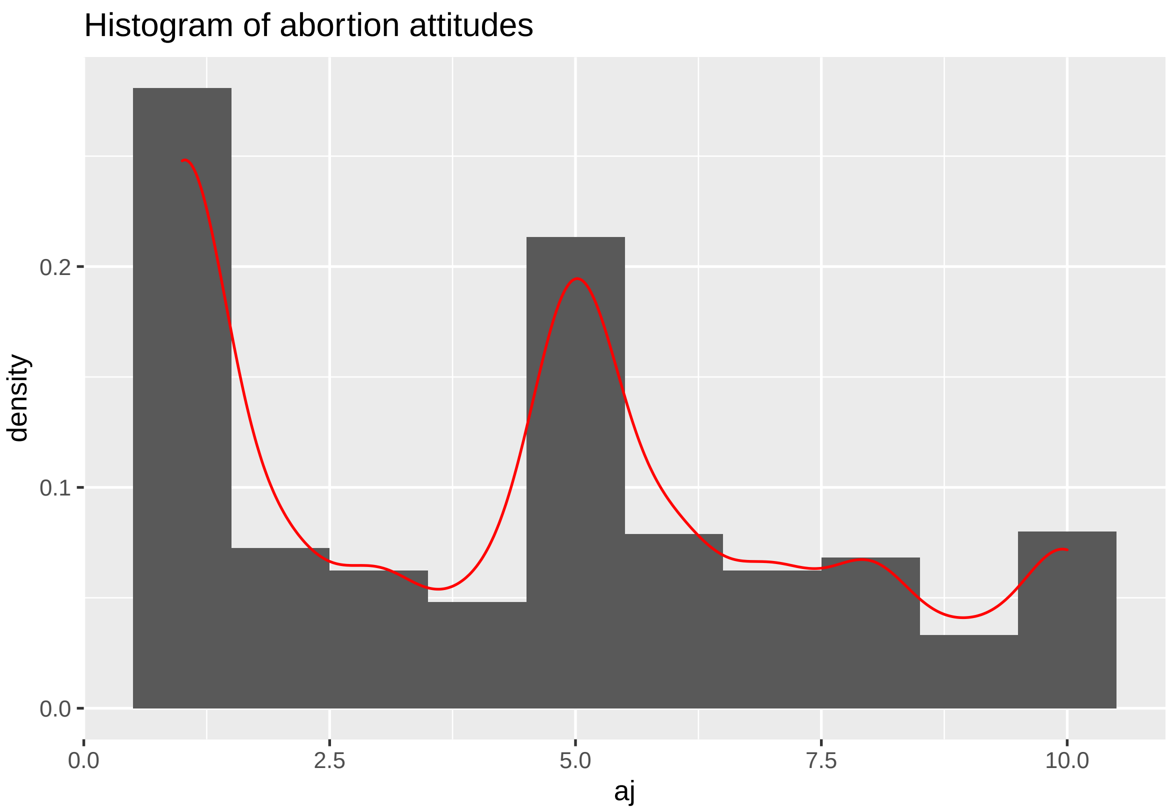
Analysis approach

Paste-updated-content-here

```
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.

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



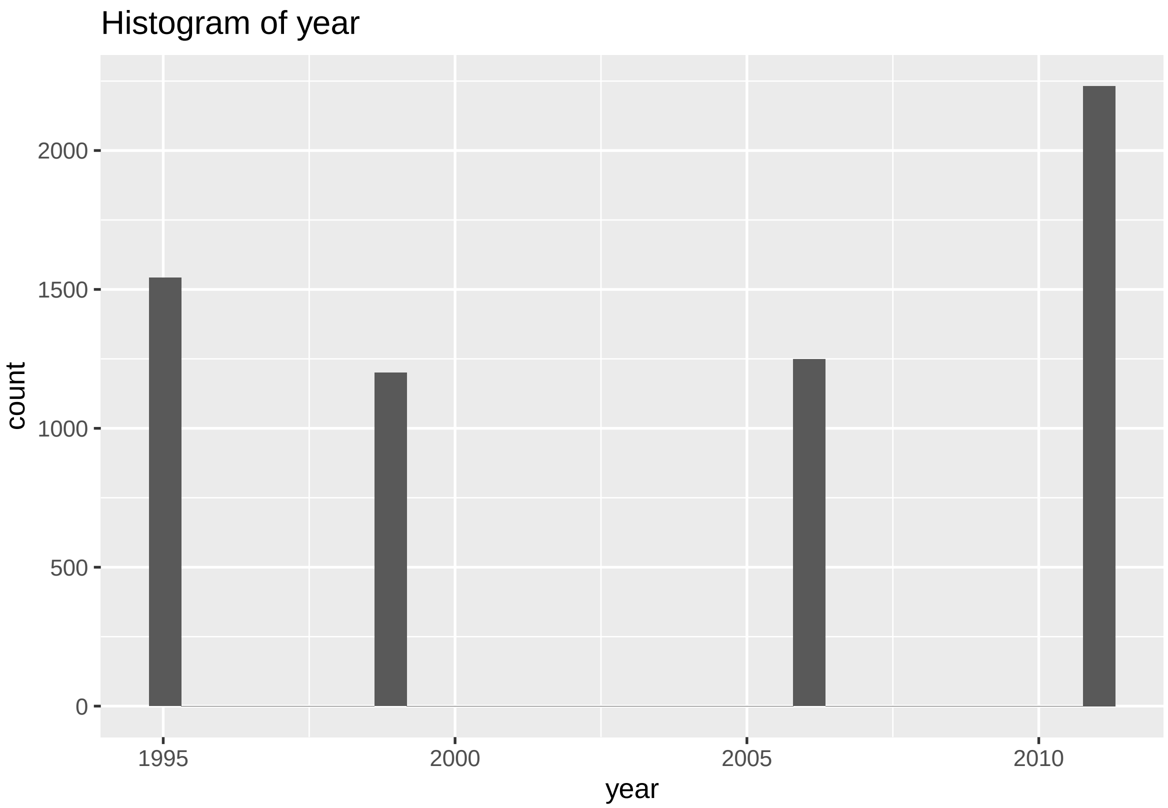
```
summary(abortion_data$aj)
```

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

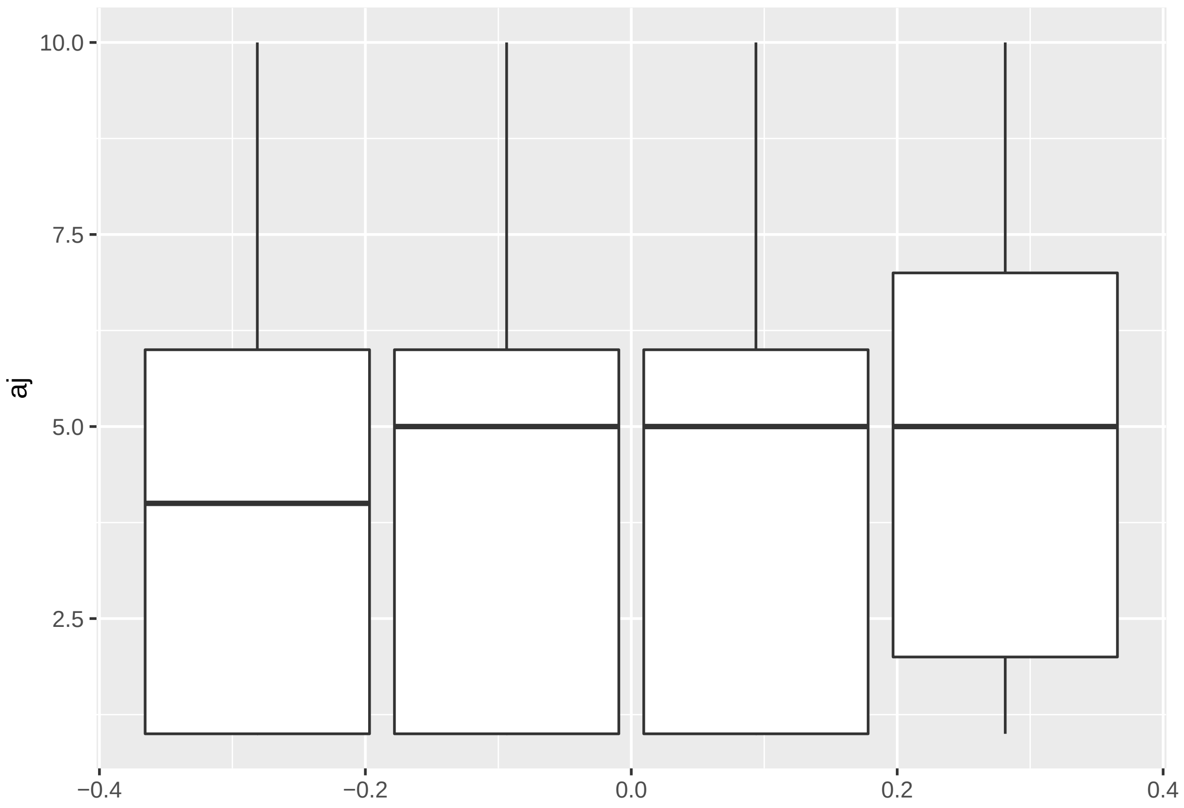
We can observe from this histogram that the distribution of the outcome variable is not a bell shape, and it is trimodal. This is likely because the question's phrasing is similar to a yes/no question, but respondents were asked to give their level of agreement on a scale of 1-10. This may result in 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 extend our exploratory data analysis (EDA) to some predictor variables of interest; namely, the year, ideology, Child Autonomy Index, Importance of God, Respect for Authority and National Pride predictors. The EDA for each variable comprises a histogram and a boxplot of the response variable grouped by predictor value. Additionally, we have a jitter plot to explore the potential for an interaction effect between year and importance of God.

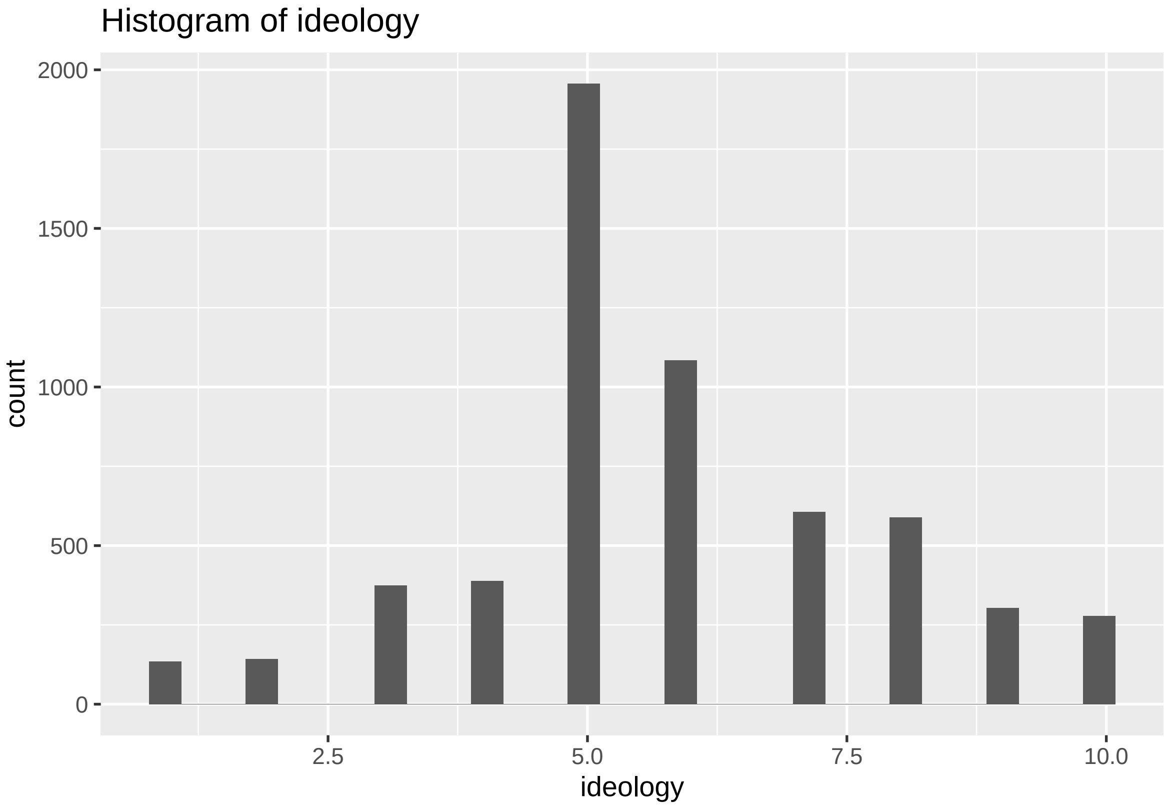
```
#Year of survey
ggplot(abortion_data, aes(x = year)) +
  geom_histogram() +
  labs(title = "Histogram of year")
```



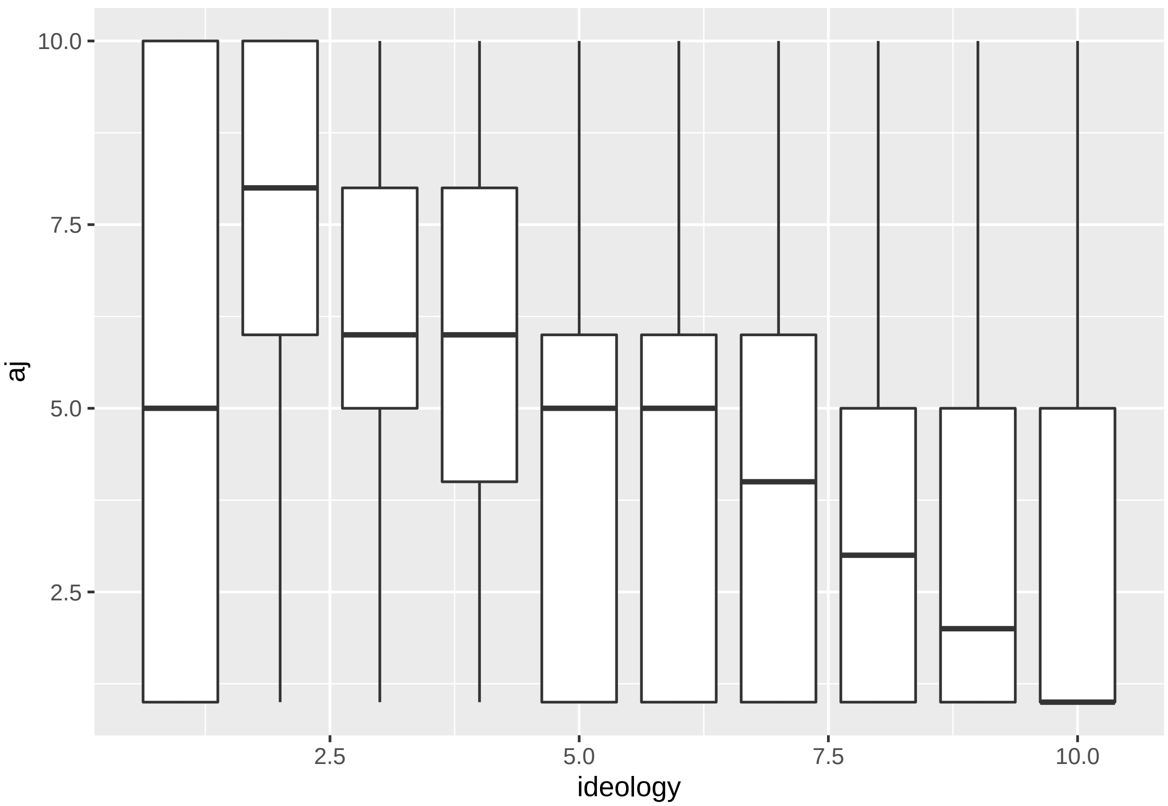
```
ggplot(abortion_data, aes(group = year, y = aj)) +
  geom_boxplot()
```



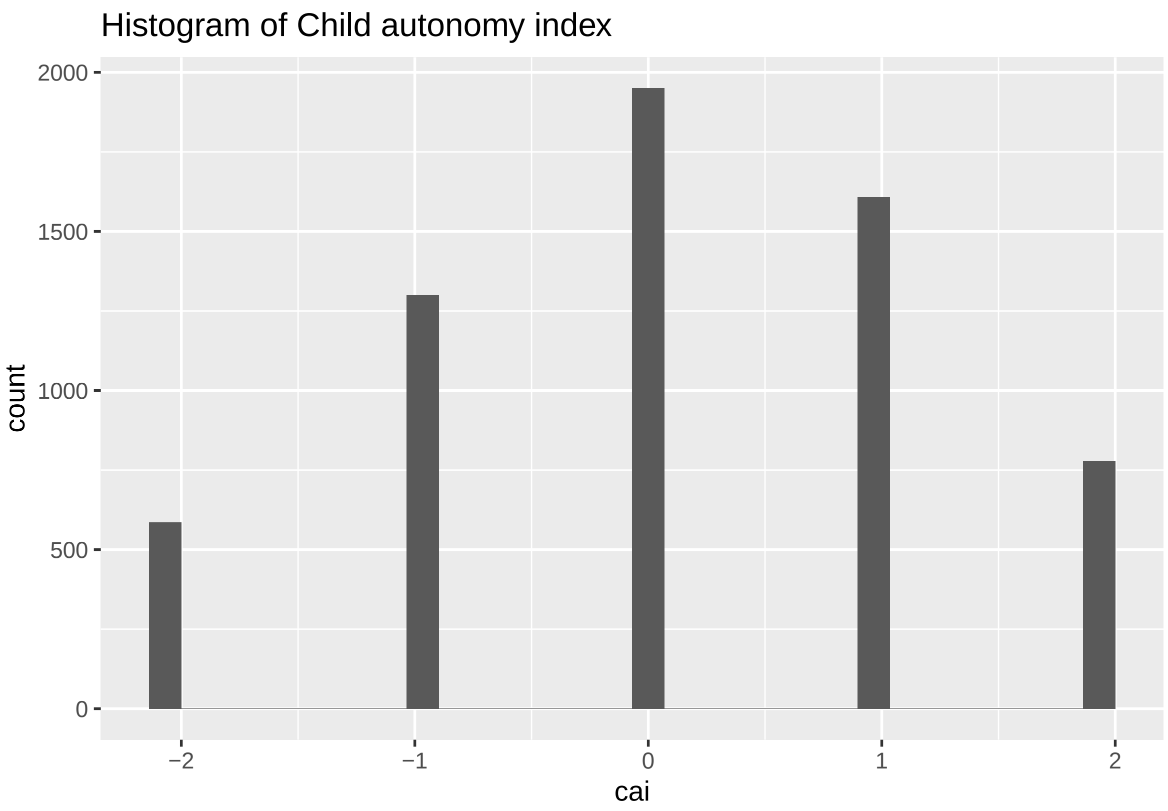
```
# ideology
ggplot(abortion_data, aes(x = ideology)) +
  geom_histogram() +
  labs(title = "Histogram of ideology")
```



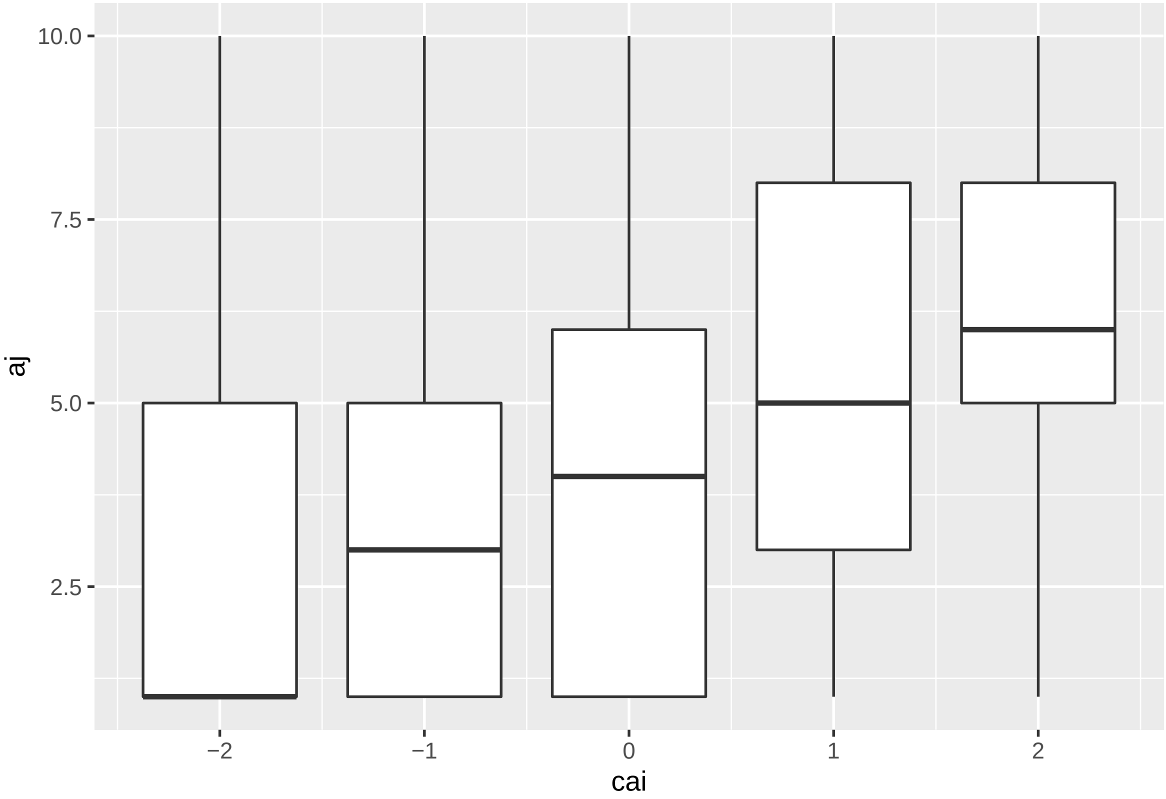
```
ggplot(abortion_data, aes(x = ideology,group = ideology, y = aj)) +
  geom_boxplot()
```



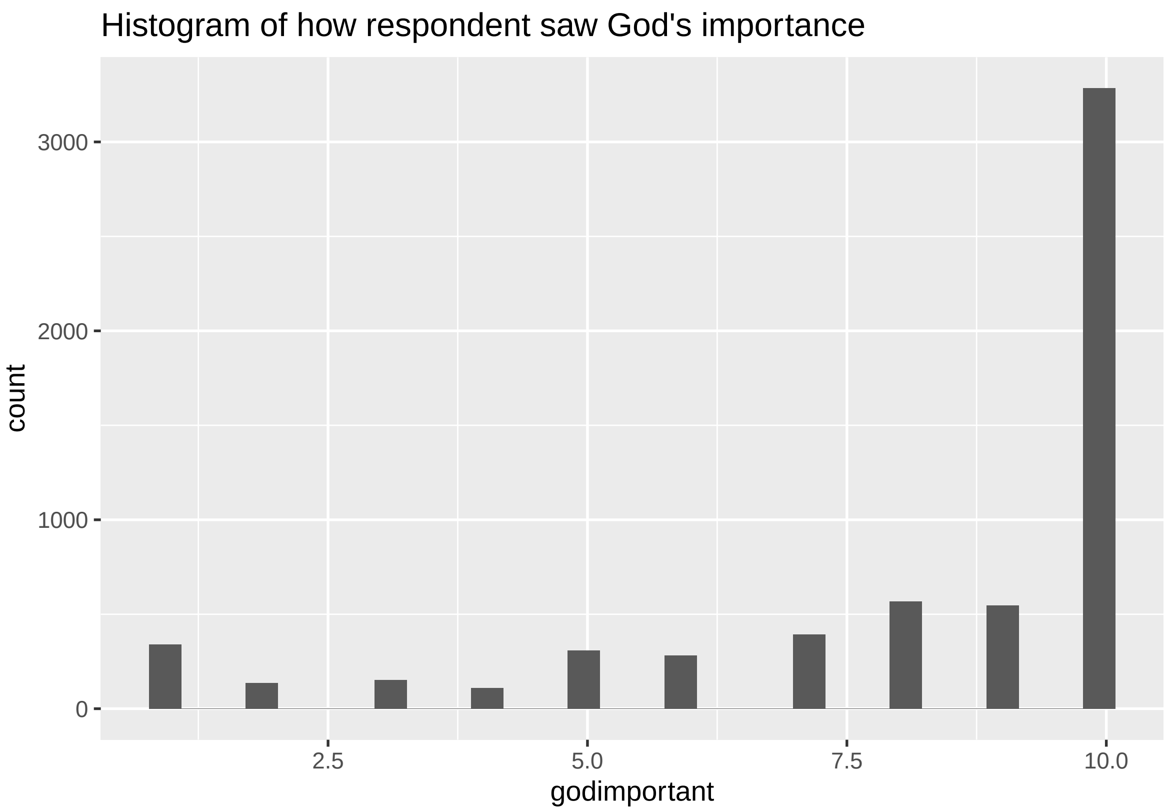
```
# Child autonomy index
ggplot(abortion_data, aes(x = cai)) +
  geom_histogram() +
  labs(title = "Histogram of Child autonomy index")
```



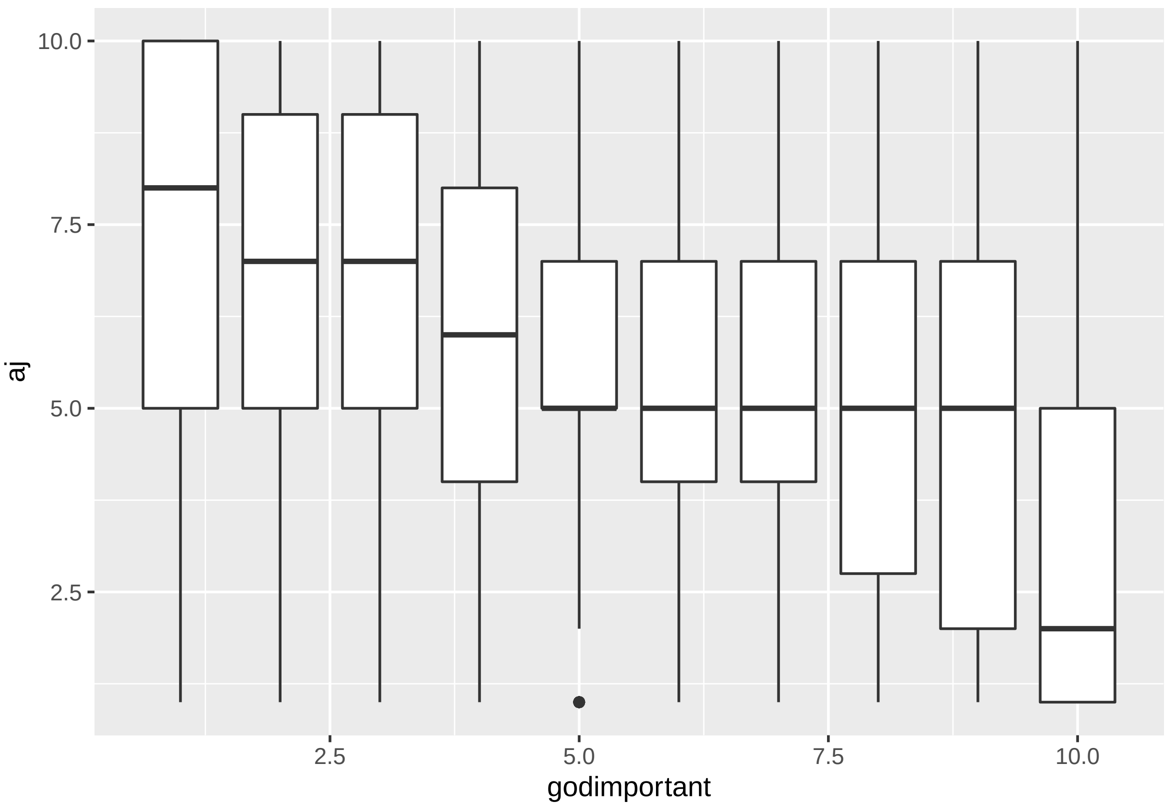
```
ggplot(abortion_data, aes(x = cai, group = cai, y = aj)) +
  geom_boxplot()
```



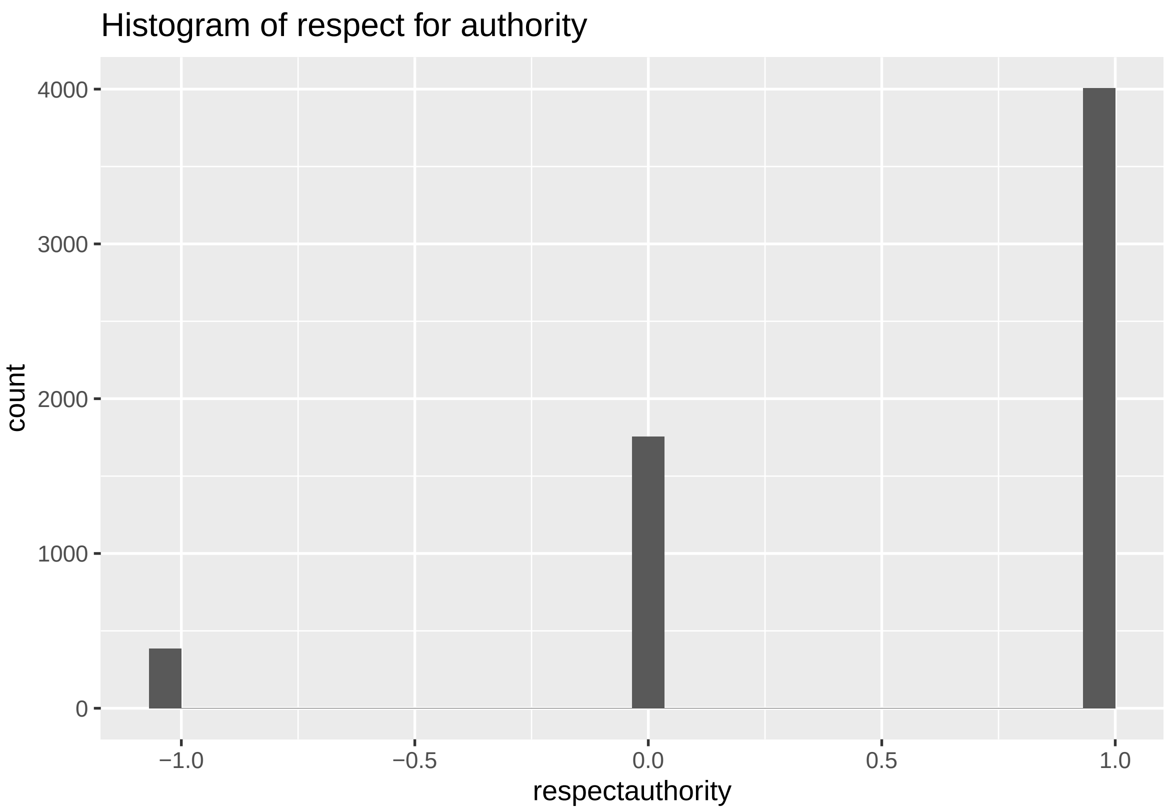
```
# Importance of God
ggplot(abortion_data, aes(x = godimportant)) +
  geom_histogram() +
  labs(title = "Histogram of how respondent saw God's importance")
```



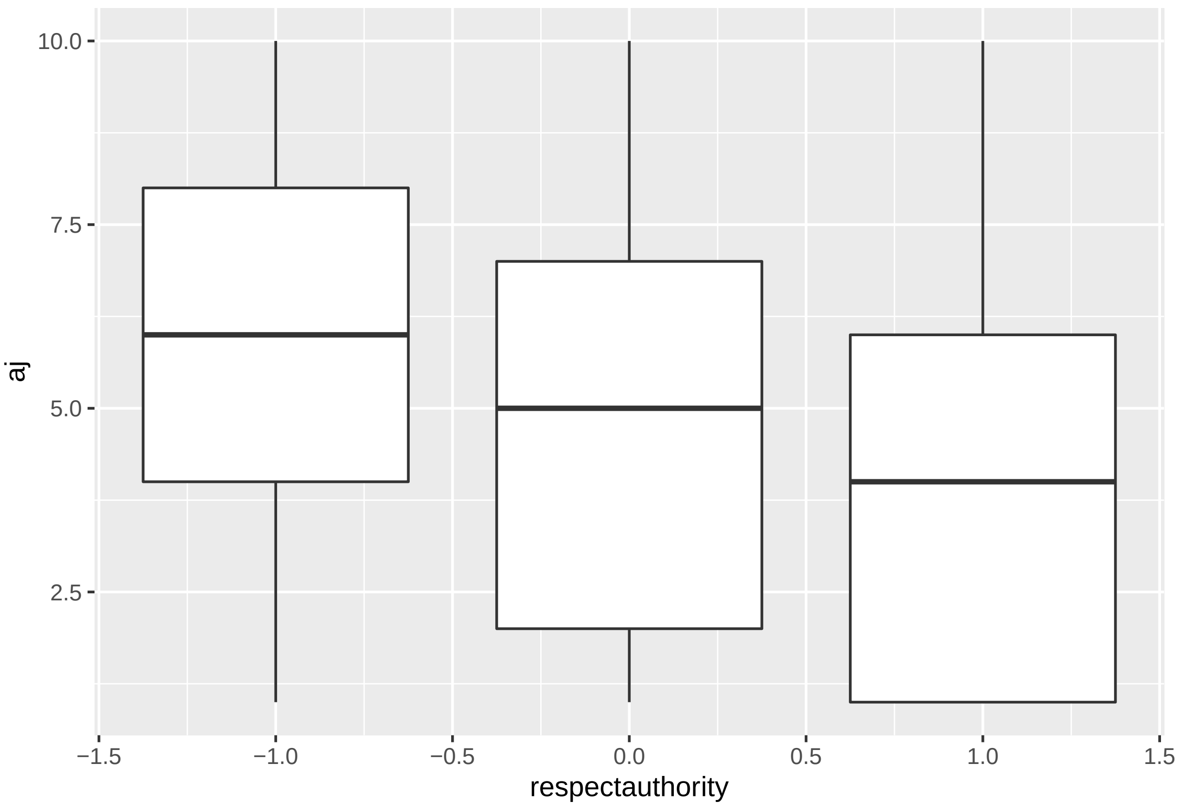
```
ggplot(abortion_data, aes(x = godimportant, group = godimportant, y = aj)) +
  geom_boxplot()
```



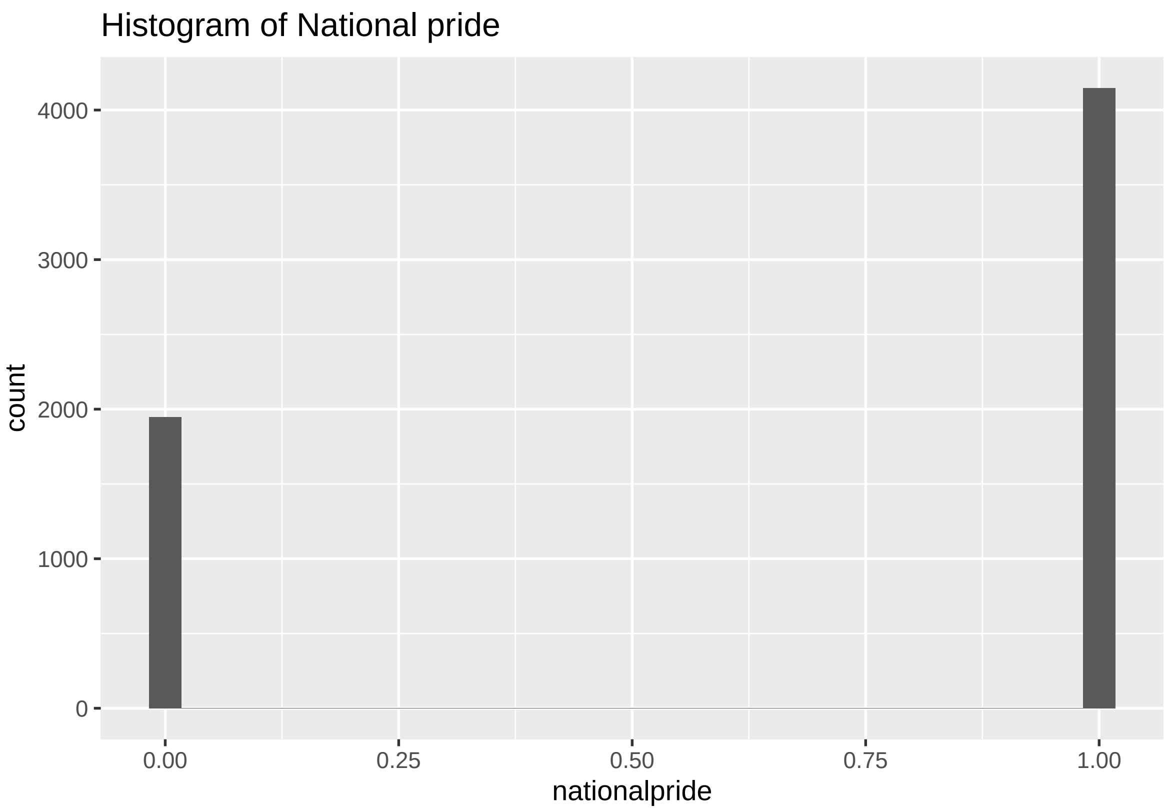
```
# Respect for authority
ggplot(abortion_data, aes(x = respectauthority)) +
  geom_histogram() +
  labs(title = "Histogram of respect for authority")
```



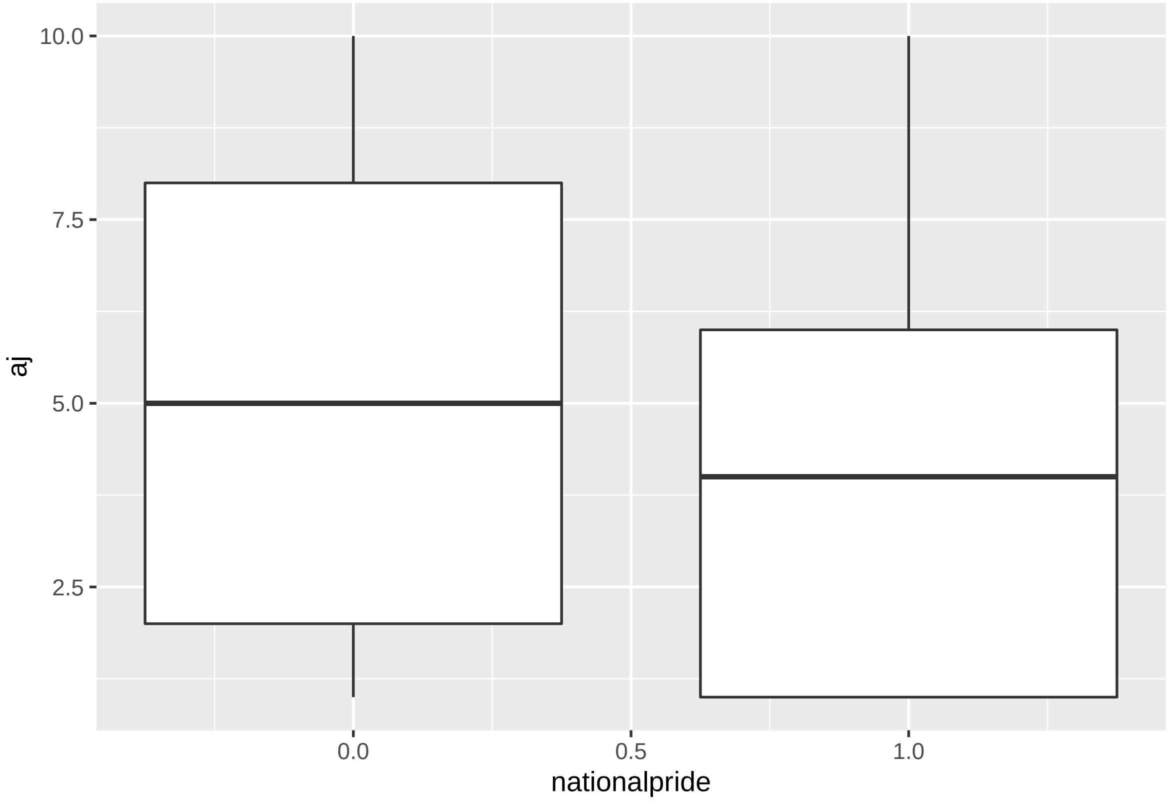
```
ggplot(abortion_data, aes(x = respectauthority, group = respectauthority, y = aj)) +
  geom_boxplot()
```



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

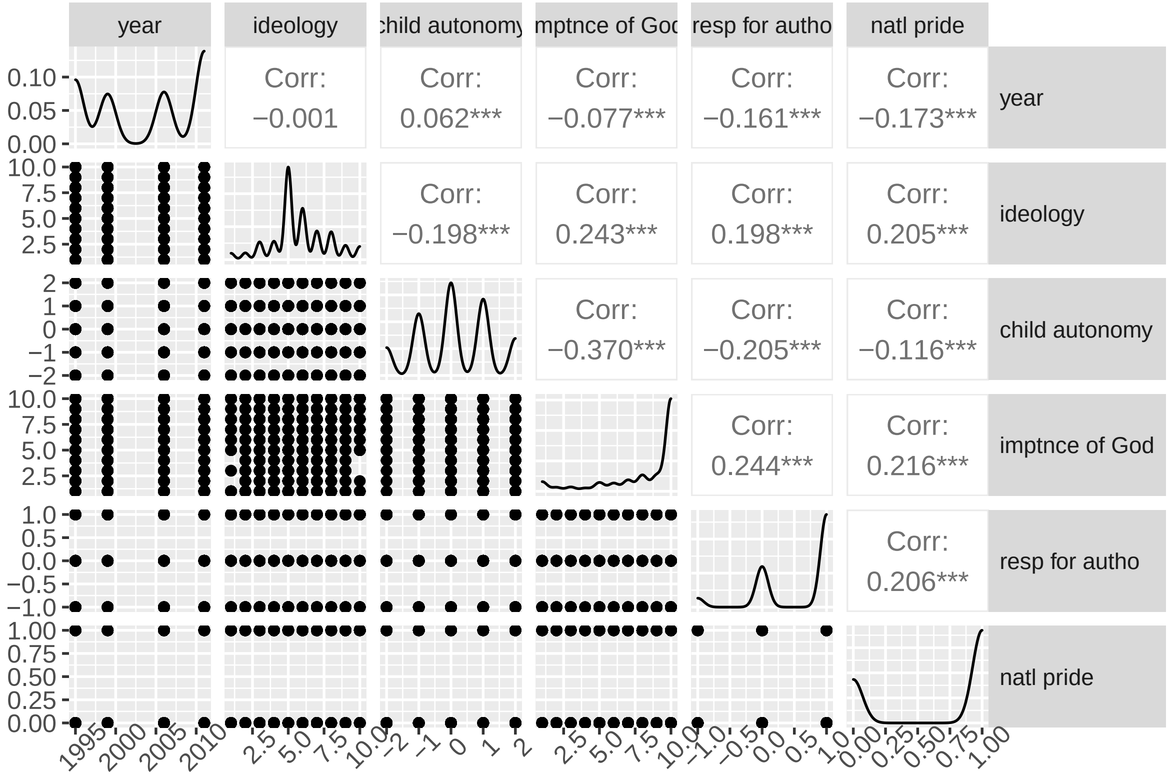


```
ggplot(abortion_data, aes(x = nationalpride, group = nationalpride, y = aj)) +
  geom_boxplot()
```



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

```r
ggpairs(abortion_data,
  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)
    )
```



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 `respectauthority` 0.206 `respectauthority` and `cai` 0.205

METHOD

We will conduct a Multiple Linear Regression (MLR) on the attitudes on the justifiability of abortion against several other predictor variables found in this data set. This is because our outcome variable, the attitude towards abortion, is measured on a numeric scale from 1 to 10, and there are multiple predictor variables that we feel are potentially correlated with it.

We will now discuss the data cleaning and predictor selection process. For observations, we removed the observations that were part of "generational waves" before 1995, because many survey questions were only added from 1995 onward, and observations before 1995 have a high number of null values.

We removed the variable "wvsccode" as it is the same for all the observations. The WVSC Code depends on the country the surveys were conducted in, and all surveys in this data set were conducted in the United States of America.

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 and ruled out variables with no visible hint of relationship.

This led to our first set of predictor variables, which we designate as Recipe1: - Year - Ideology - Child Autonomy Index - Importance of God - Respect for Authority - National Pride

We also wanted to choose predictor variables that are not strongly correlated with each other, to avoid multicollinearity. From our EDA, we found that several predictors had correlations of 0.2 or higher with each other. Given these correlations, we decided to make a second recipe (`Recipe2`) with the following predictor variables, after excluding those with a high correlation with the `godimportant` variable:

- Year
- Importance of God

Finally, we considered that the distribution for the `godimportant` variable is heavily left-skewed in our data set, which we observed in the EDA. This may reduce the ability of a model that relies on the "importance of God" predictor to explain the variations we observe in the outcome. The same can be said of the `respectauthority` variable, with more than 8 times the observations indicating "1" compared to the observations indicating "-1".

Consequently, our third recipe (Recipe3) excludes `godimportant` and uses the other variables that are not correlated with each other more than 0.200:

- Year
- Ideology
- Child Autonomy Index
- National Pride

Data Split

Fit Model

Note: make sure to create recipes, specify engine, make workflows, and fit models here

Inference

##RESULTS

Model Evaluation

Model Comparison

Model Diagnosis

Discussion of Influential Points

Data dictionary

The data dictionary can be found here.