Proposal

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-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 movement 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

1

that policy experts are able to create and 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. 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

[Kat Cottrell]

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.

```
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")
```

Histogram of abortion attitudes
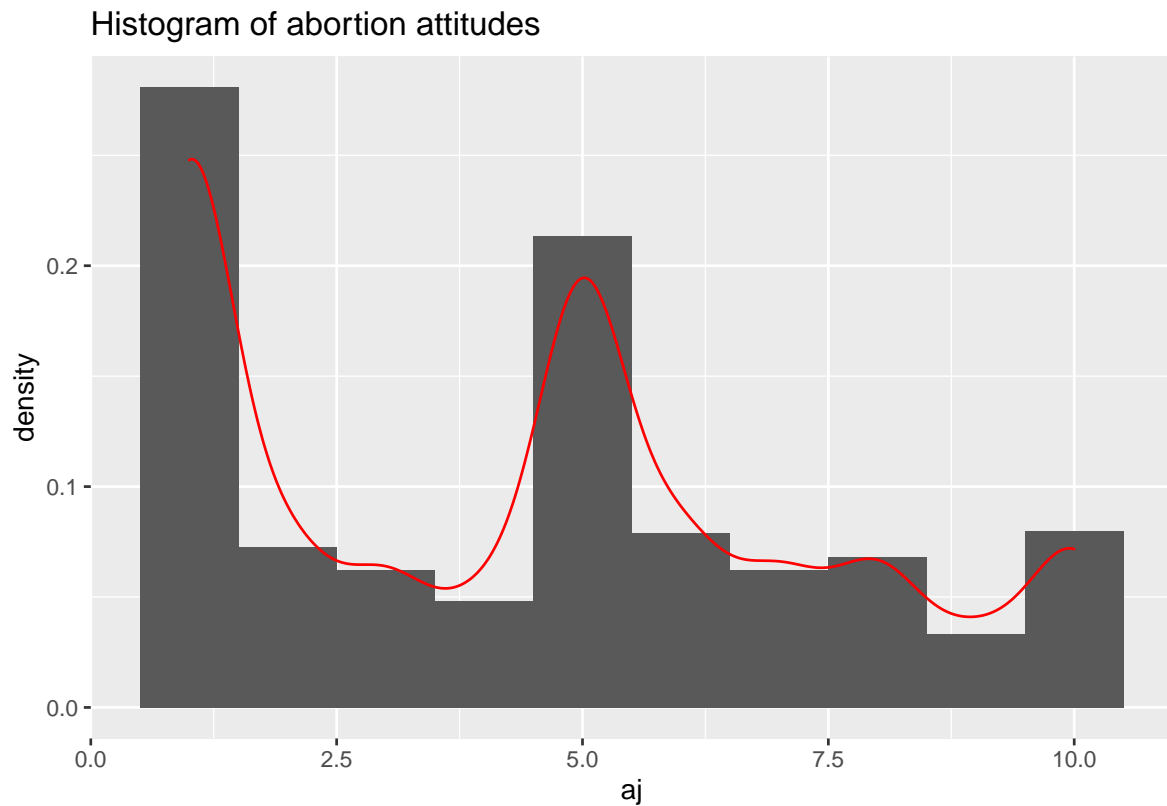

```
summary(abortion_data$aj)
```

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

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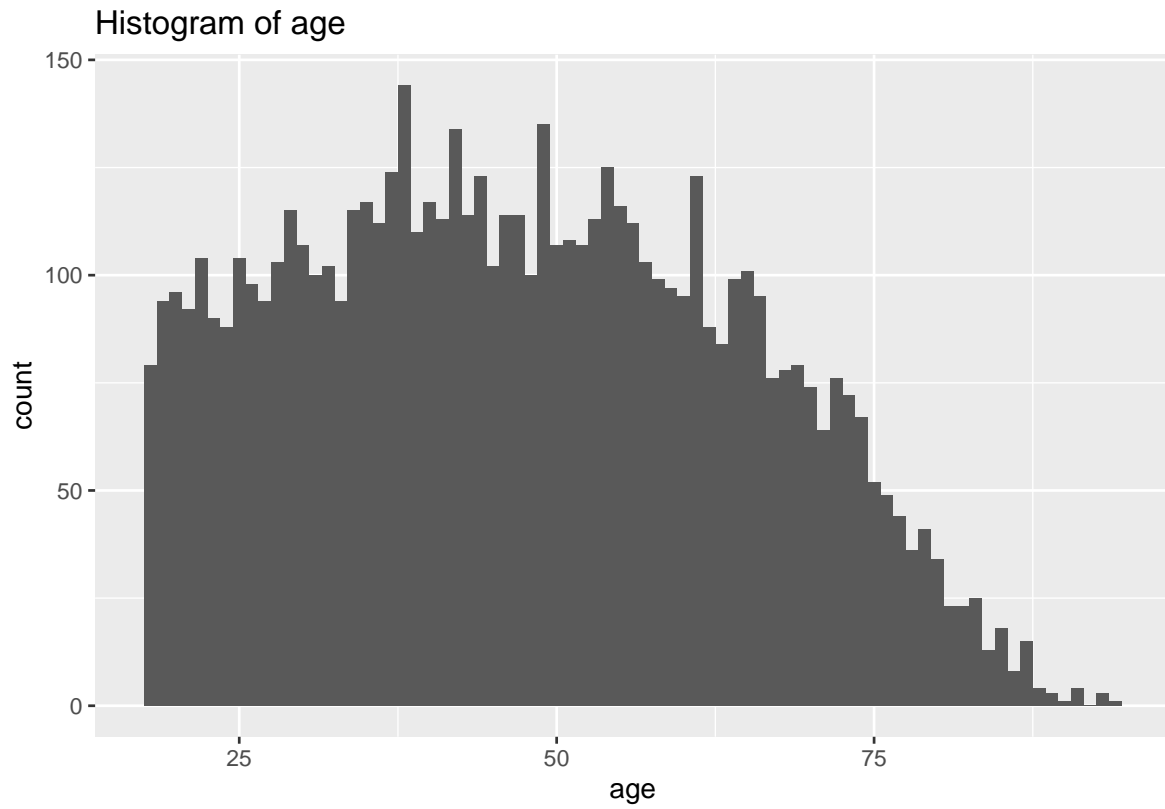
```
#Year of survey
ggplot(abortion_data, aes(x = year)) +
  geom_histogram() +
  labs(title = "Histogram of year")
```
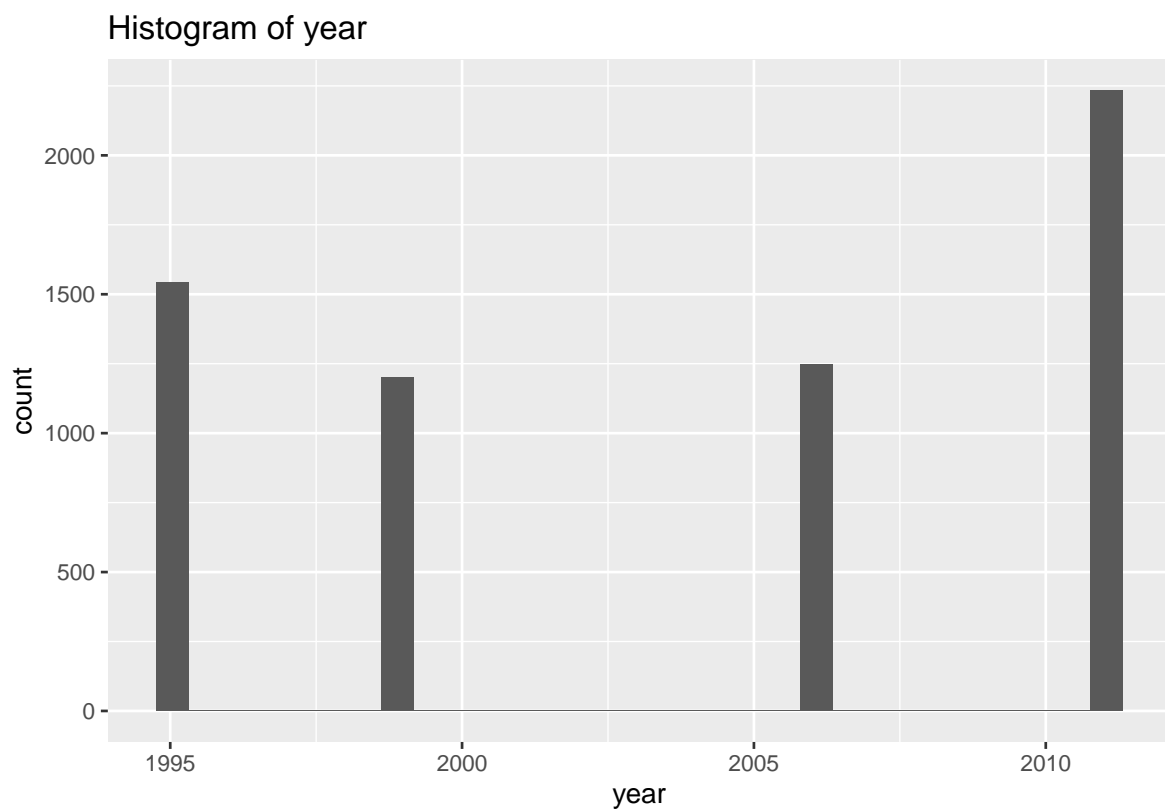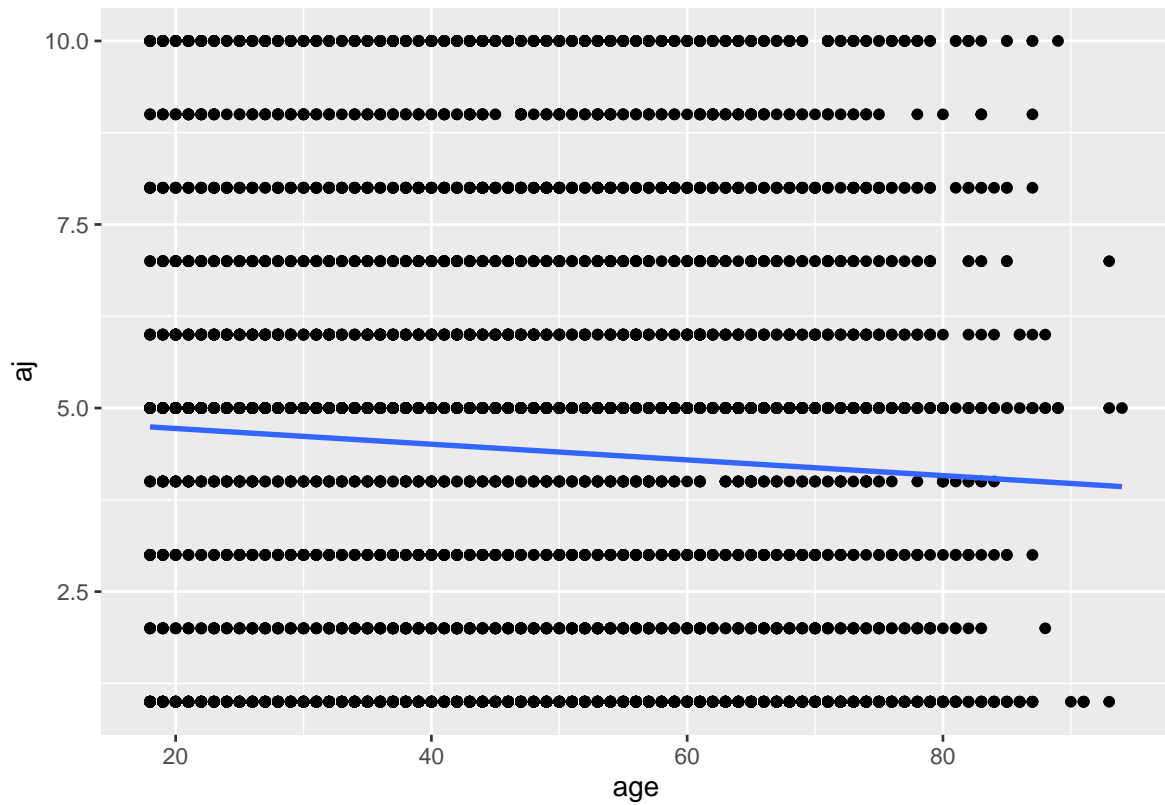
Histogram of year


```
ggplot(abortion_data, aes(x = year, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
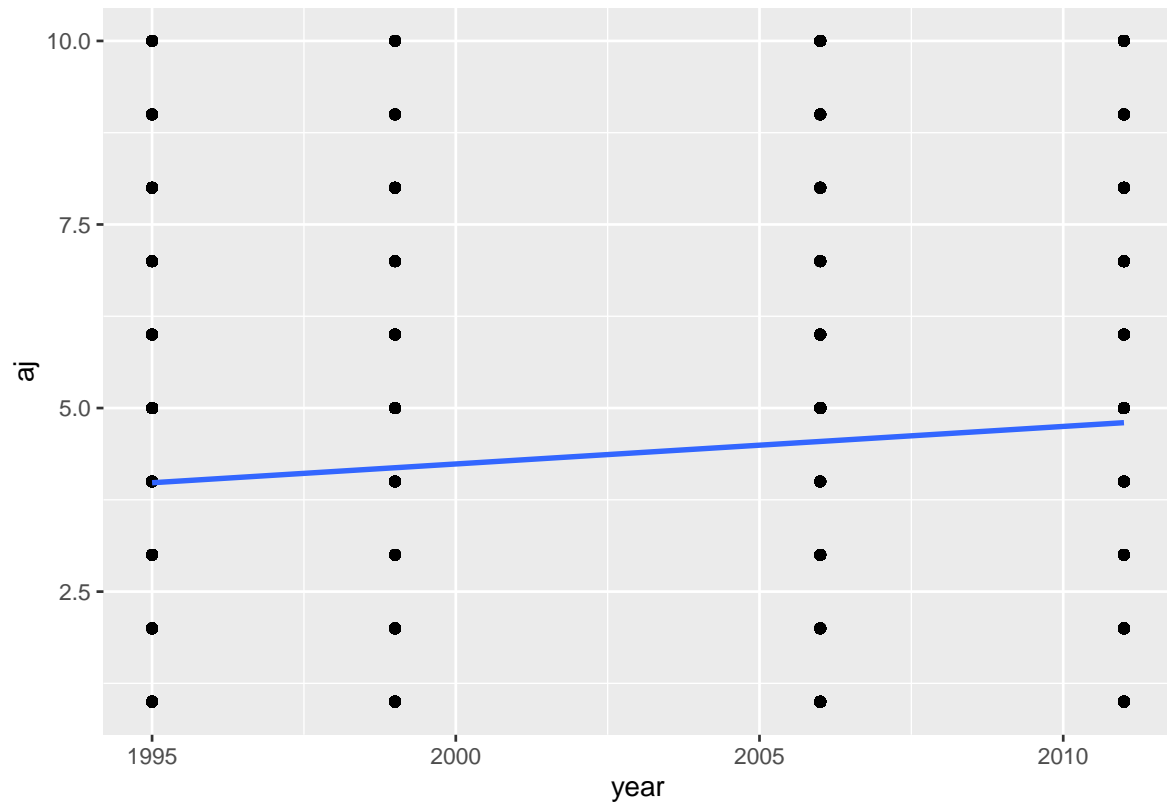
```
#Age
ggplot(abortion_data, aes(x = age)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Histogram of age")
```
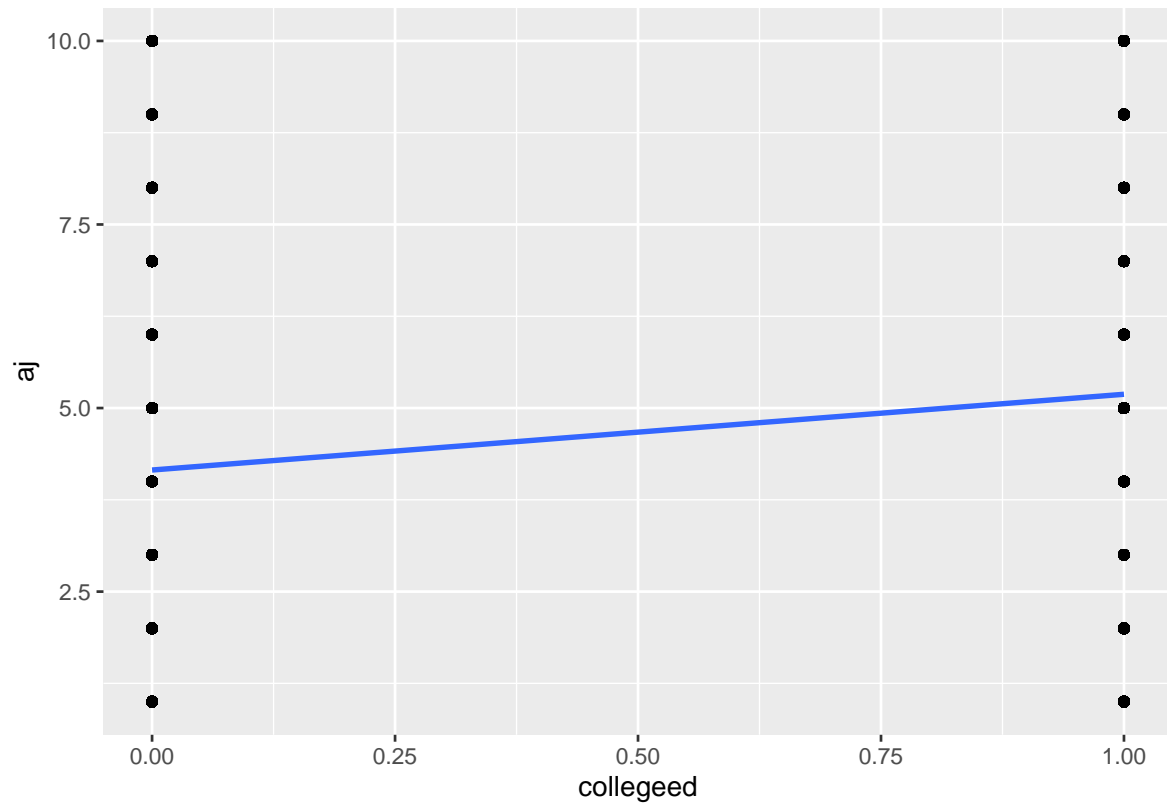
Histogram of age

```
ggplot(abortion_data, aes(x = age, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
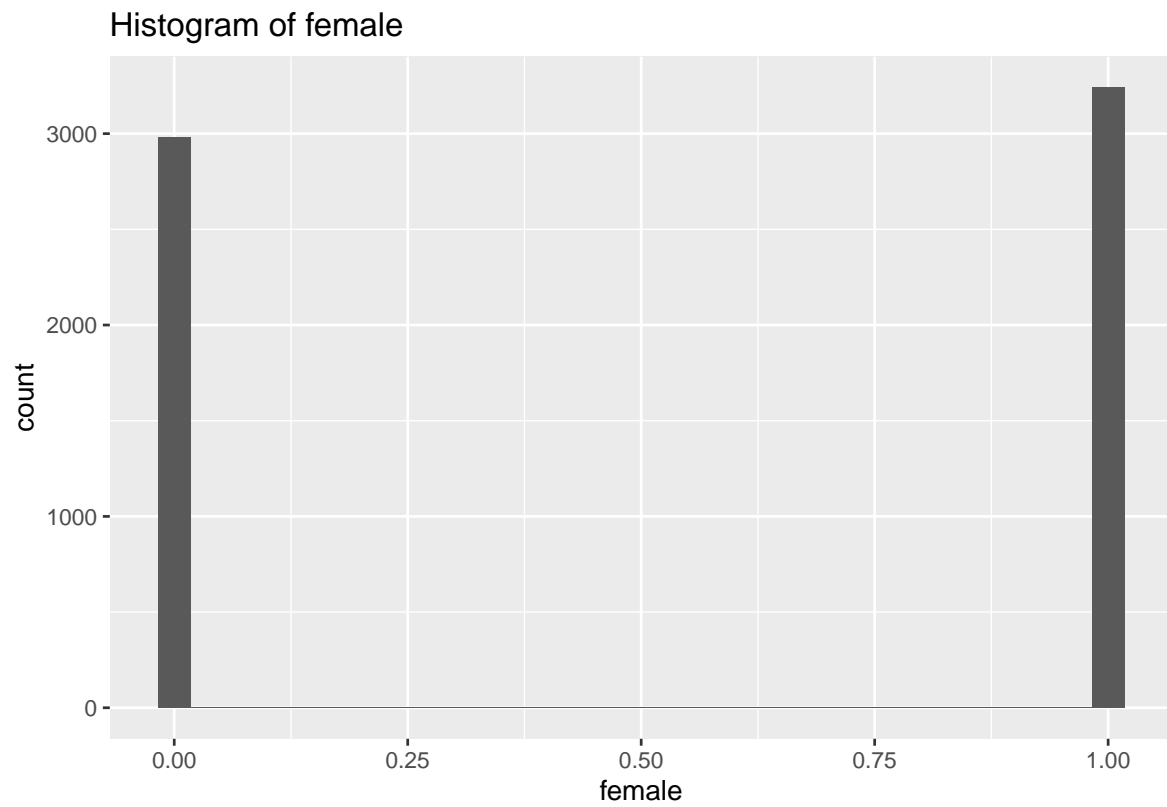
```
# College education
ggplot(abortion_data, aes(x = collegeed)) +
  geom_histogram() +
  labs(title = "Histogram of College education")
```
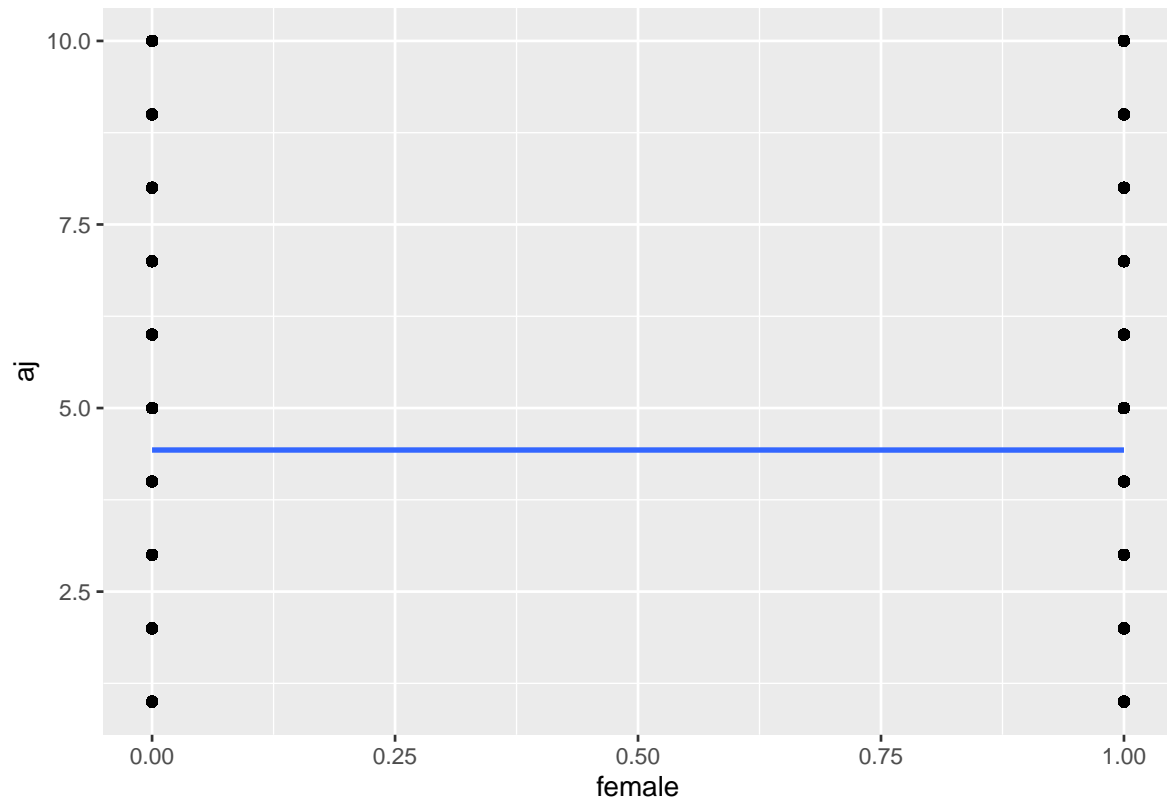
Histogram of College education
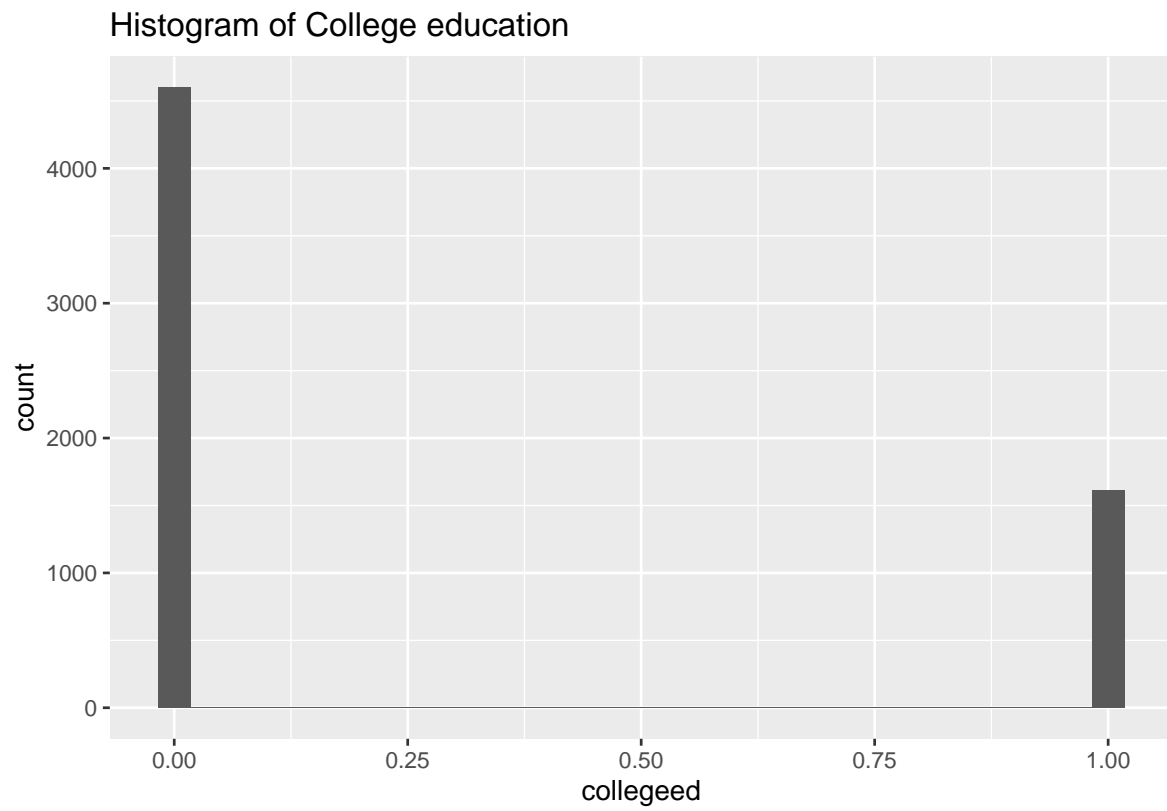

```
ggplot(abortion_data, aes(x = collegeed, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
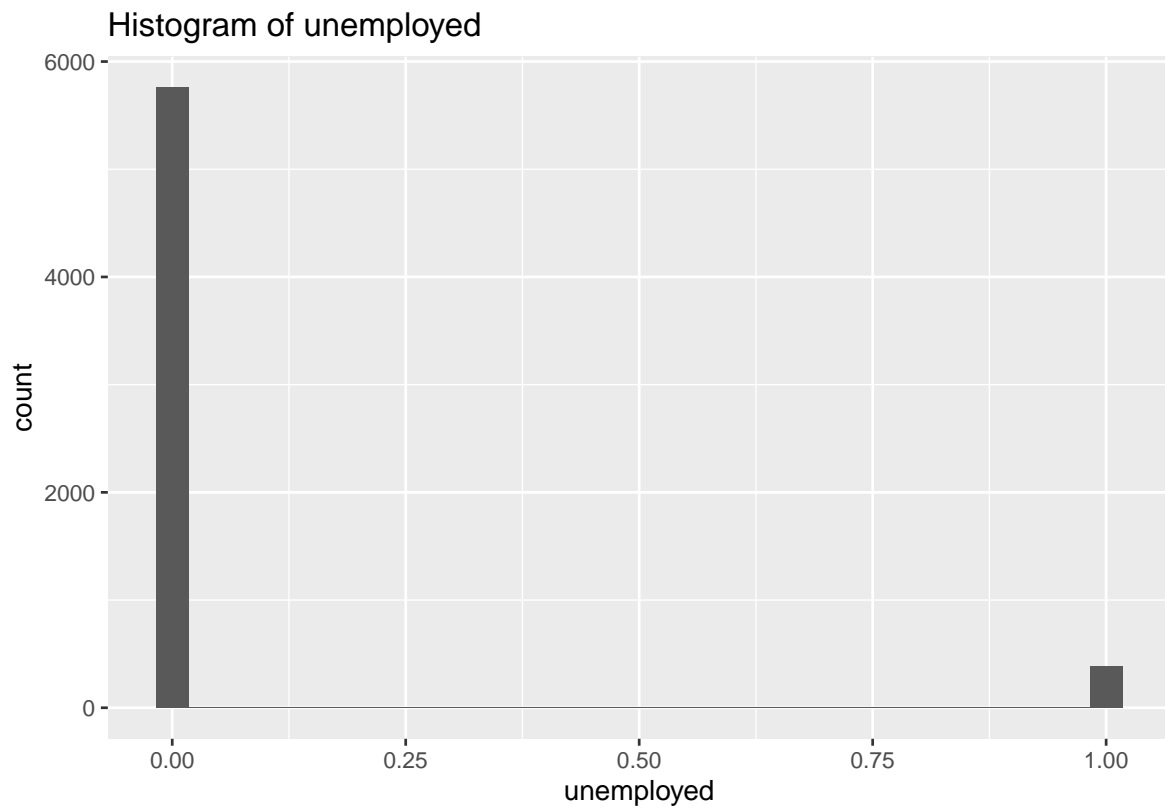
```
# Female
ggplot(abortion_data, aes(x = female)) +
  geom_histogram() +
  labs(title = "Histogram of female")
```
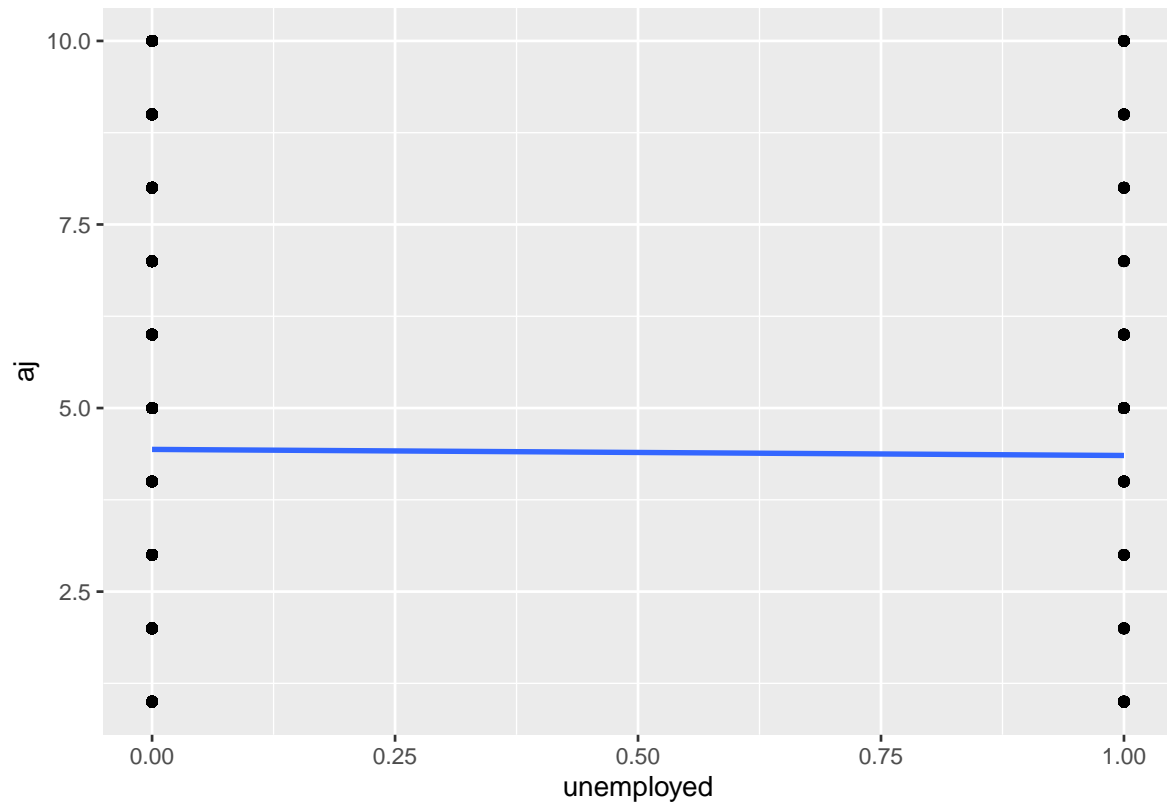
Histogram of female


```
ggplot(abortion_data, aes(x = female, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
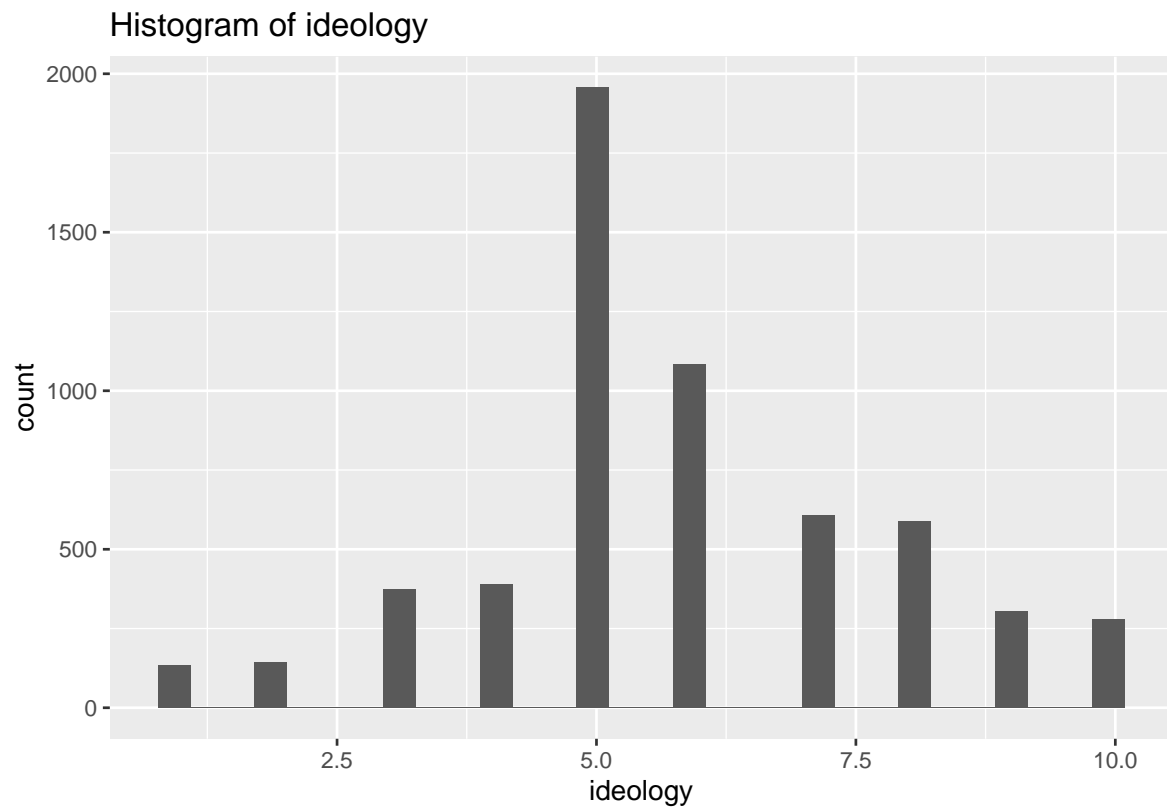
```
# Unemployed
ggplot(abortion_data, aes(x = unemployed)) +
  geom_histogram() +
  labs(title = "Histogram of unemployed")
```
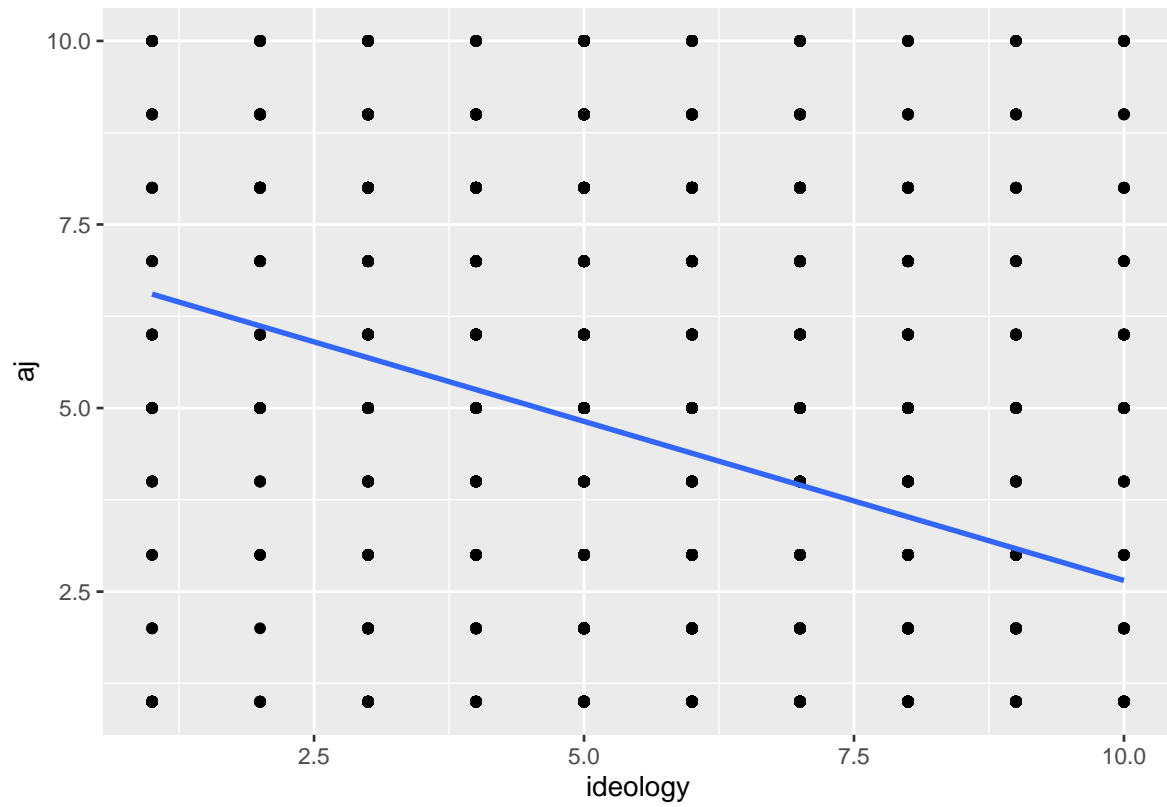
Histogram of unemployed


```
ggplot(abortion_data, aes(x = unemployed, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

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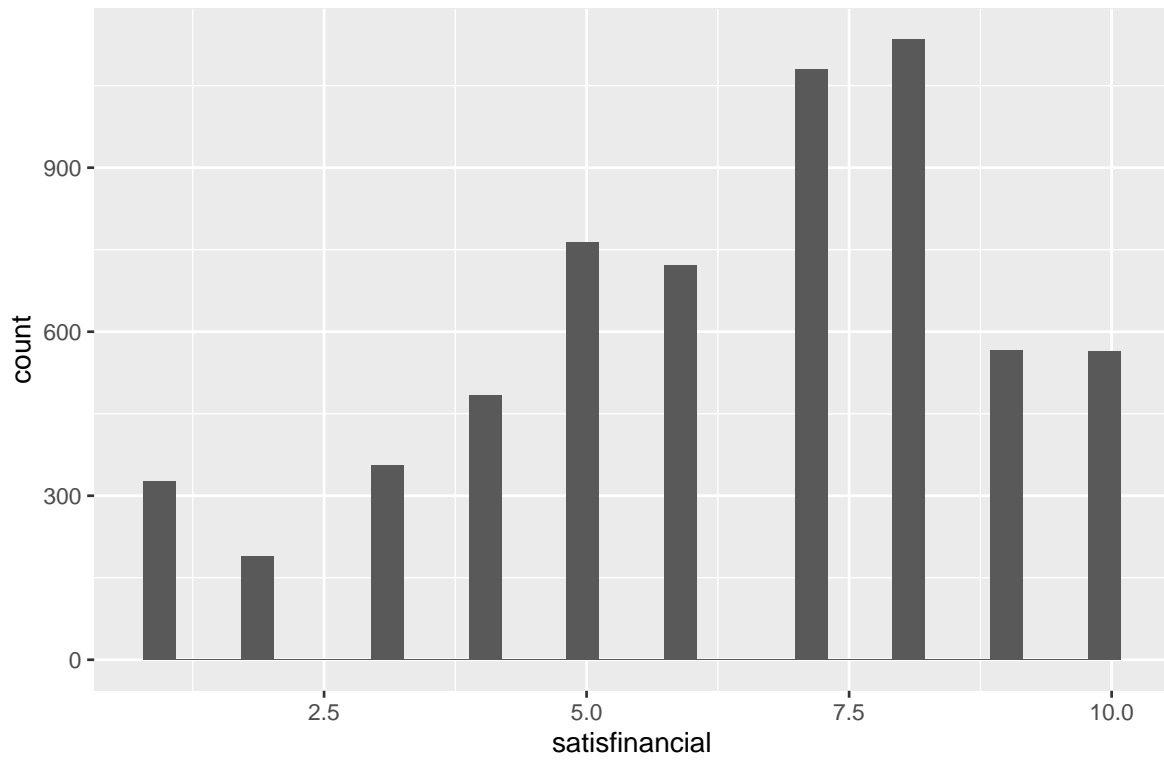
Histogram of ideology

```
ggplot(abortion_data, aes(x = ideology, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
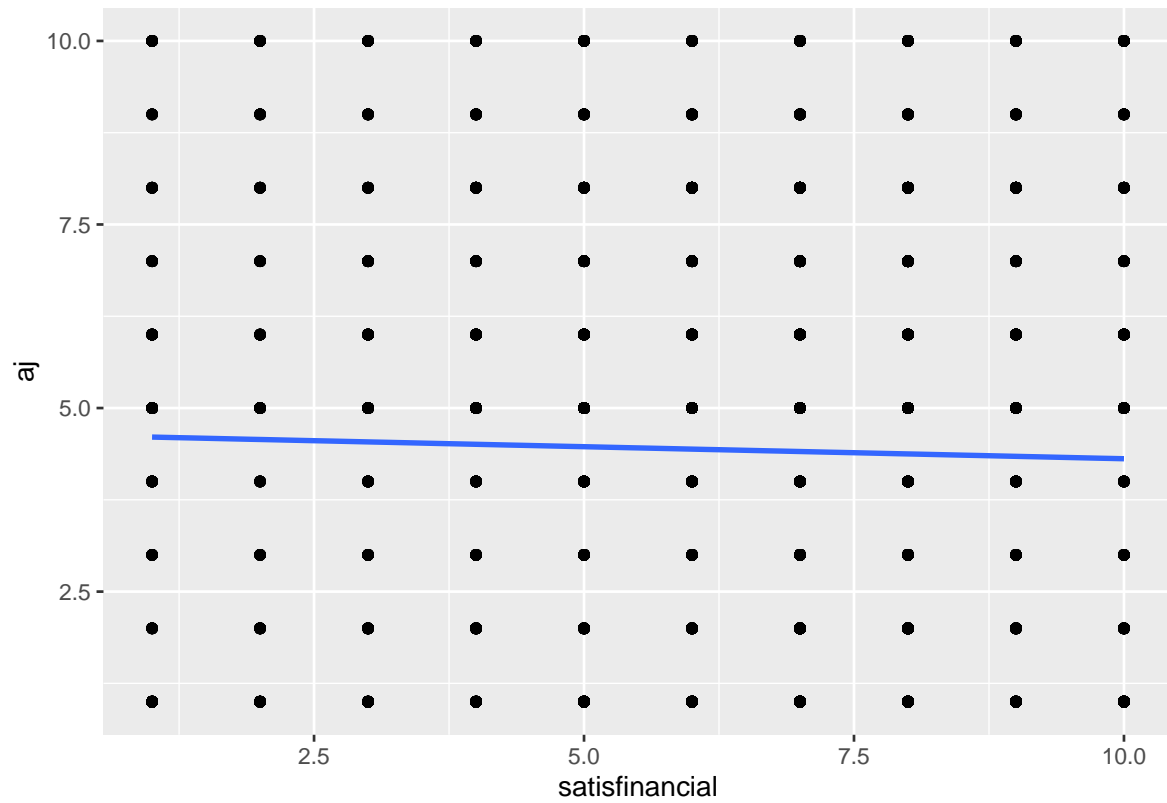
```
# Financial satisfaction
ggplot(abortion_data, aes(x = satisfinancial)) +
  geom_histogram() +
  labs(title = "Histogram of Financial satisfaction")
```
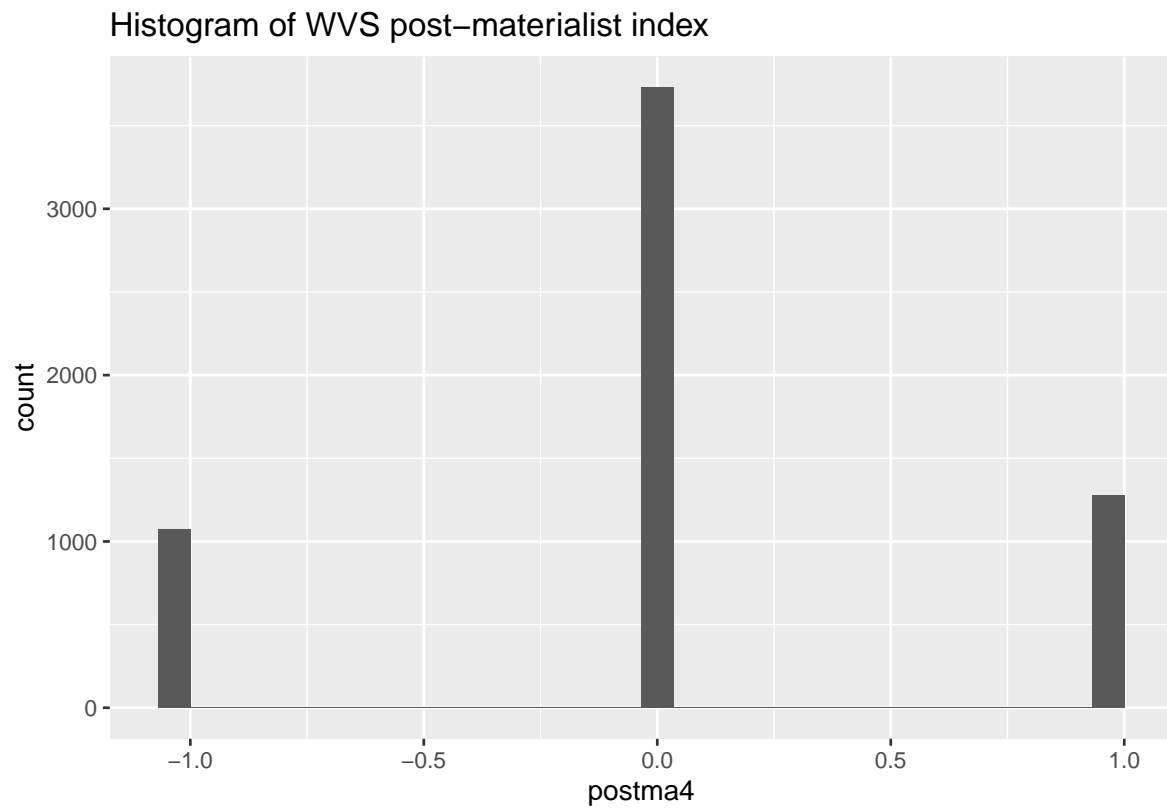
## Histogram of Financial satisfaction


```
ggplot(abortion_data, aes(x = satisfinancial, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

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

Histogram of WVS post−materialist index


```
ggplot(abortion_data, aes(x = postma4, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
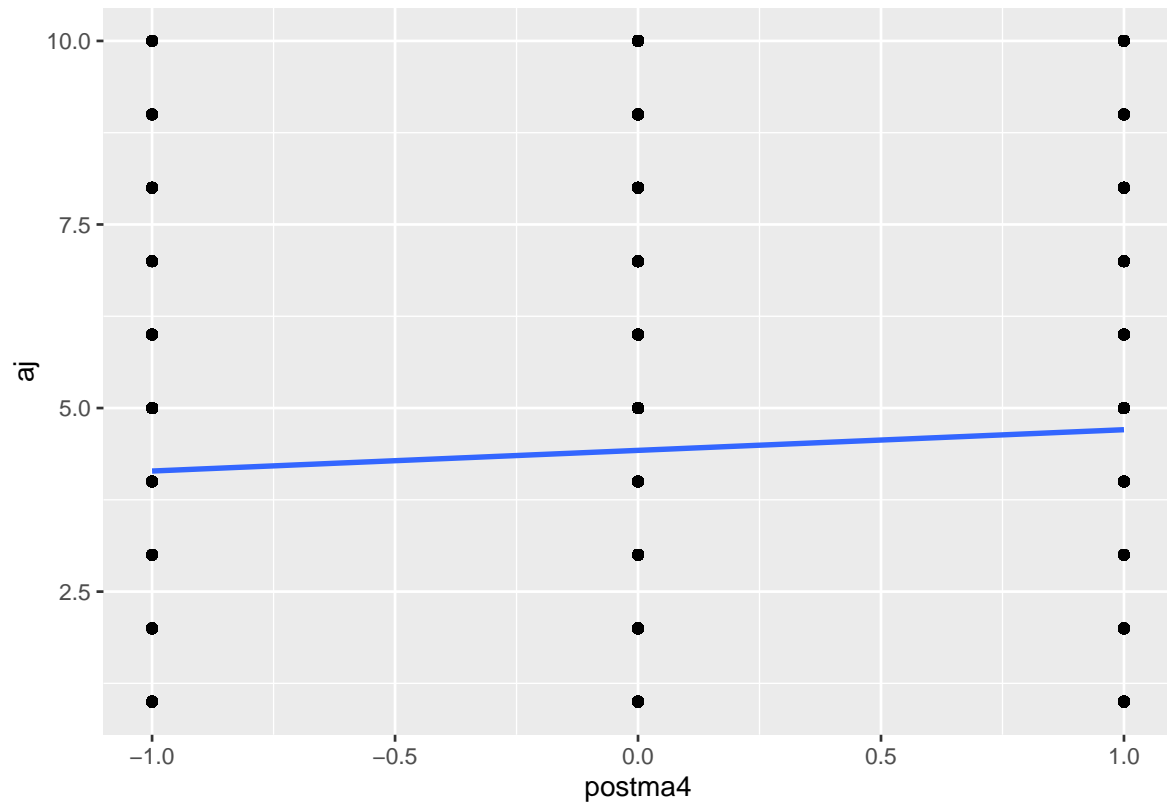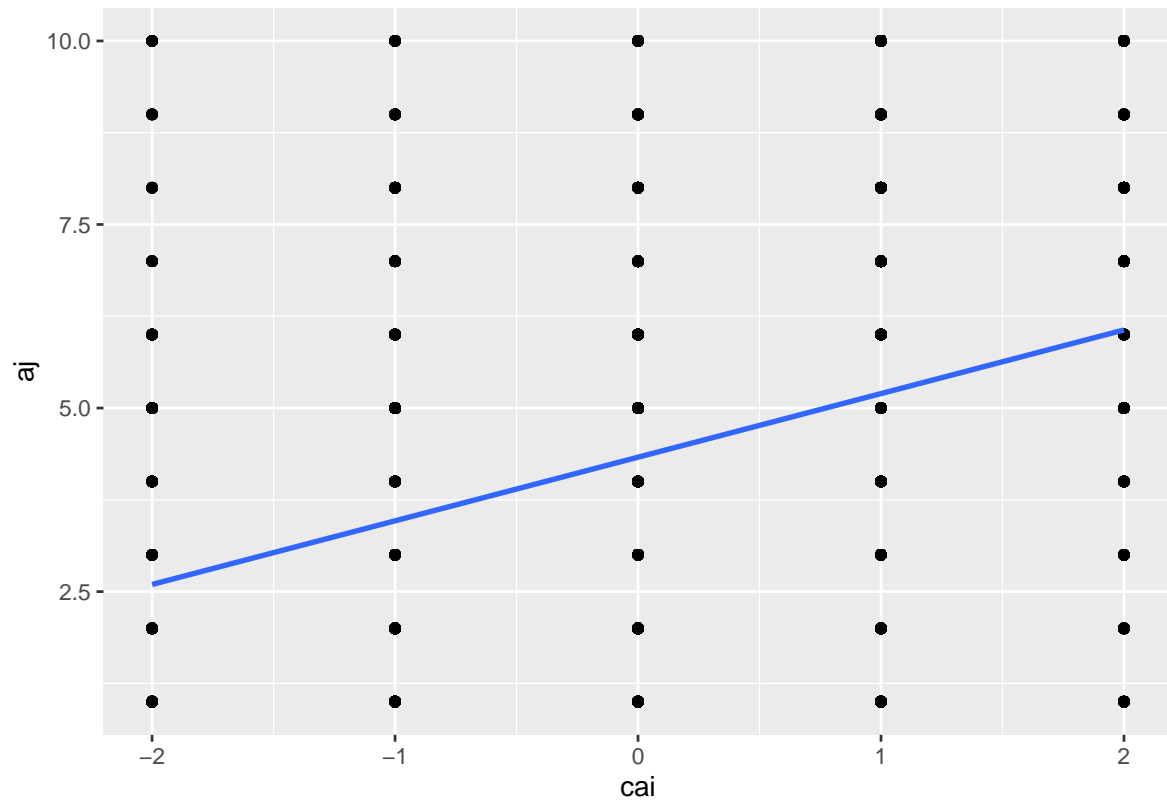
```
# Child autonomy index
ggplot(abortion_data, aes(x = cai)) +
  geom_histogram() +
  labs(title = "Histogram of Child autonomy index")
```

Histogram of Child autonomy index
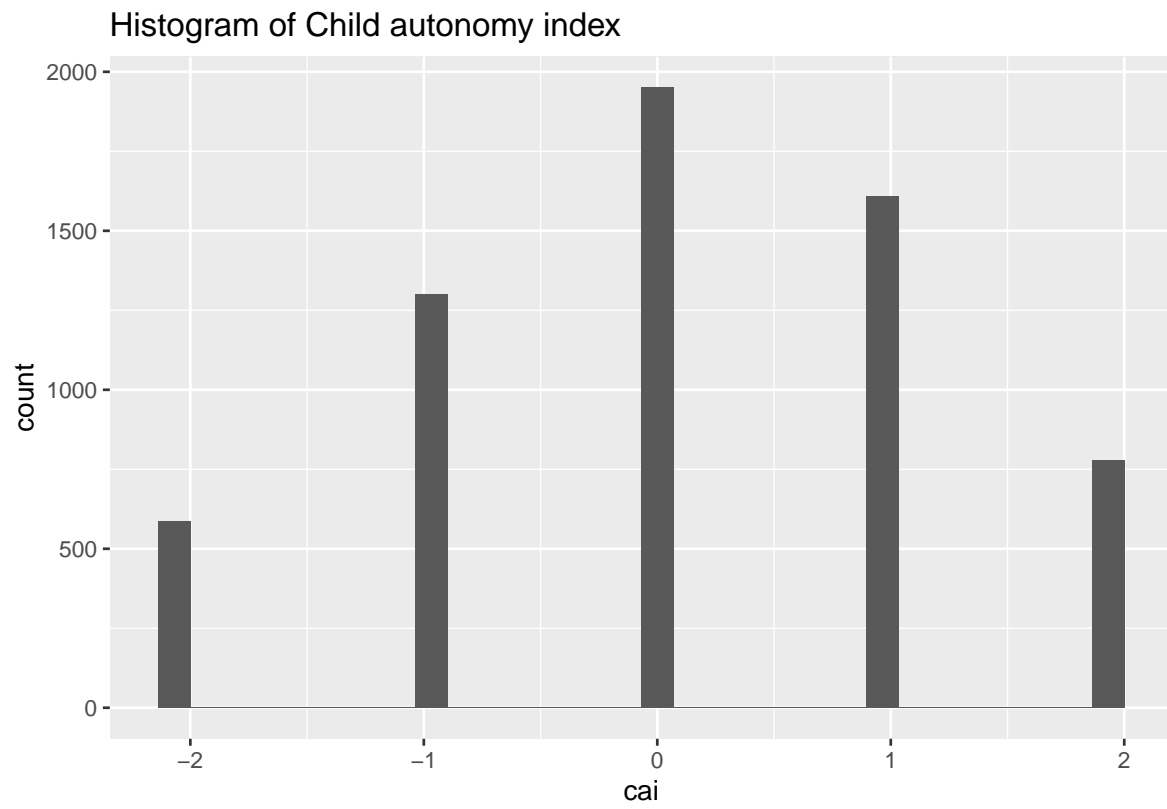

```
ggplot(abortion_data, aes(x = cai, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

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

Histogram of whether respondent trusted most people
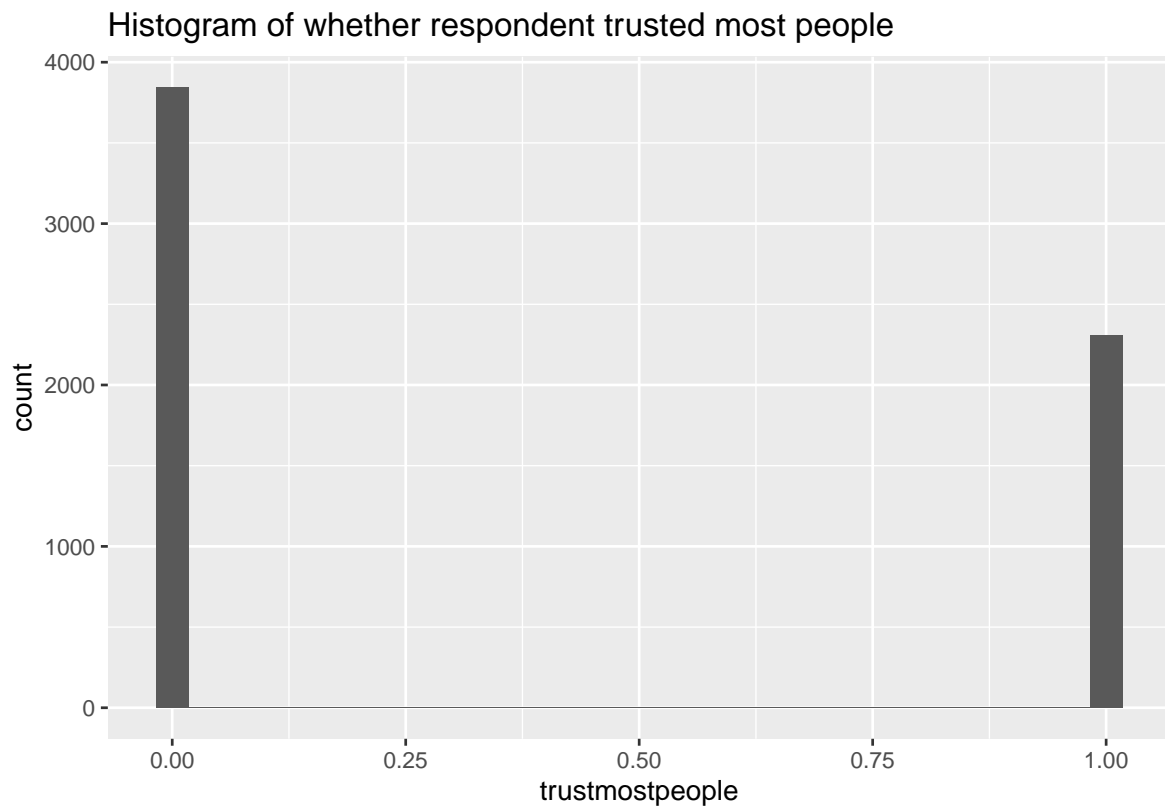
```
ggplot(abortion_data, aes(x = trustmostpeople, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
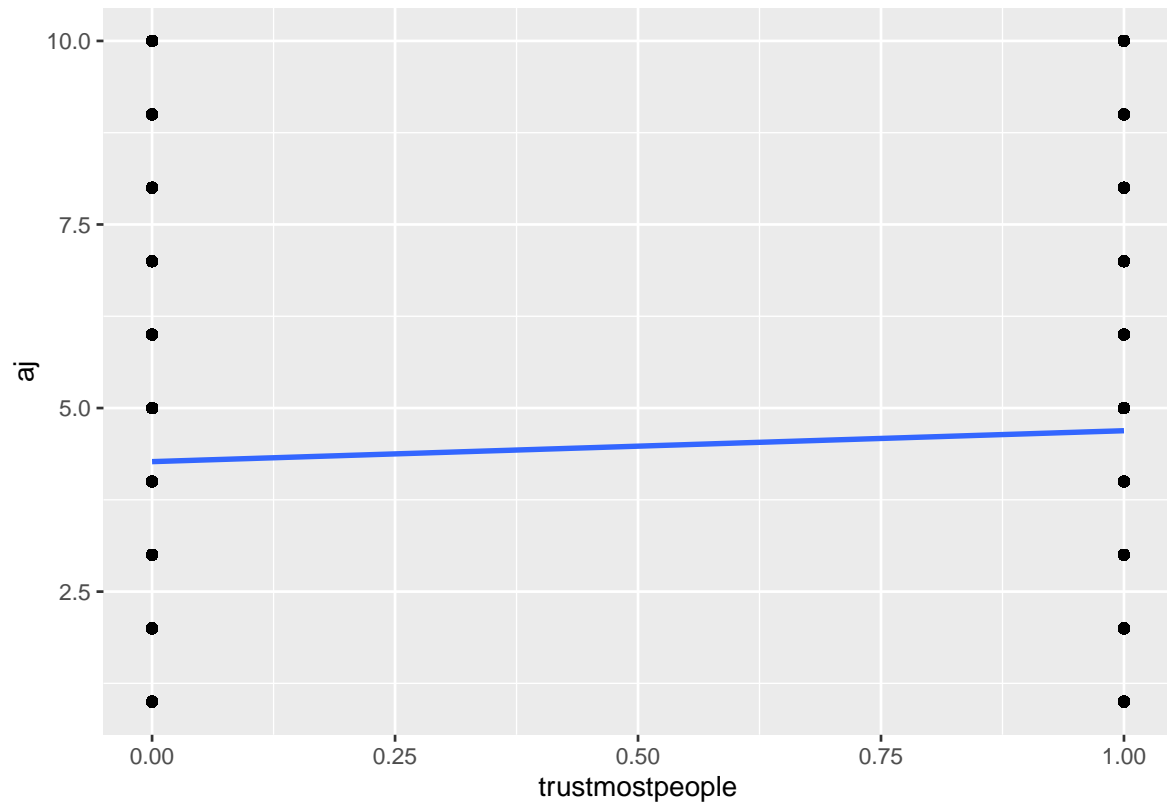
```
# Importance of God
ggplot(abortion_data, aes(x = godimportant)) +
  geom_histogram() +
  labs(title = "Histogram of how respondent saw God's importance")
```

Histogram of how respondent saw God's importance


```
ggplot(abortion_data, aes(x = godimportant, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
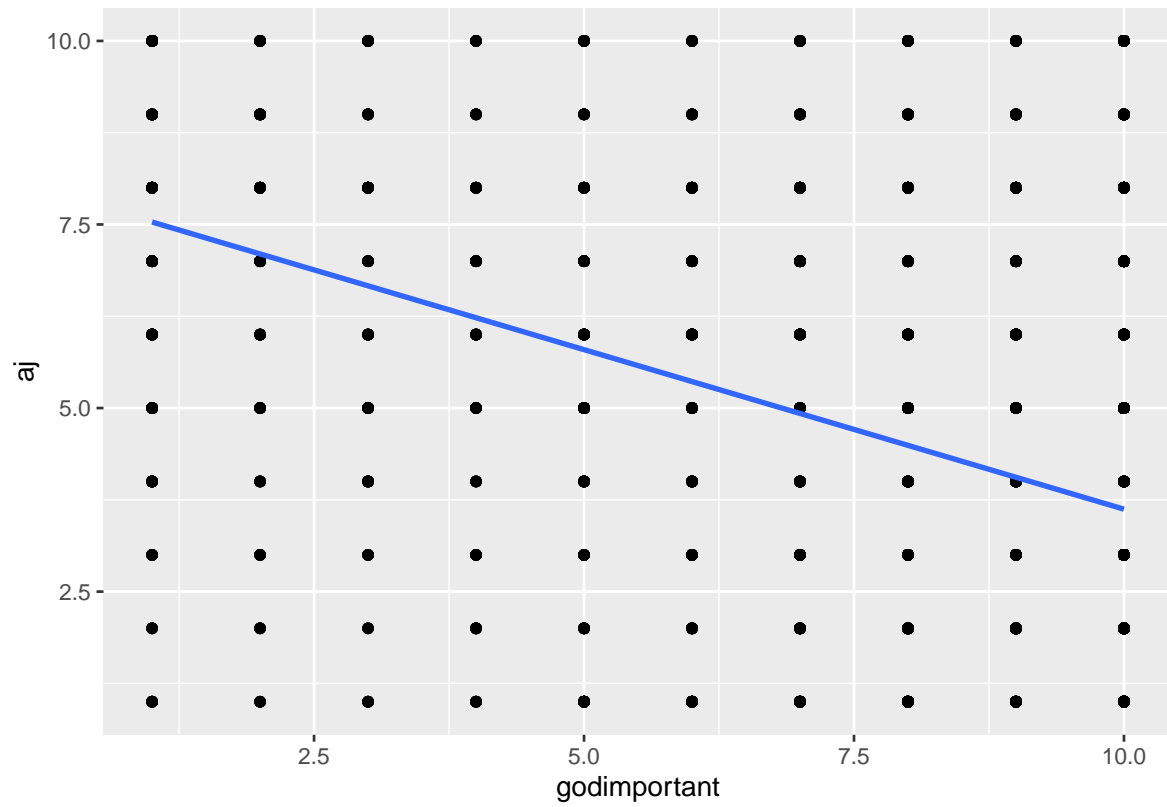
```
# Respect for authority
ggplot(abortion_data, aes(x = respectauthority)) +
  geom_histogram() +
  labs(title = "Histogram of respect for authority")
```
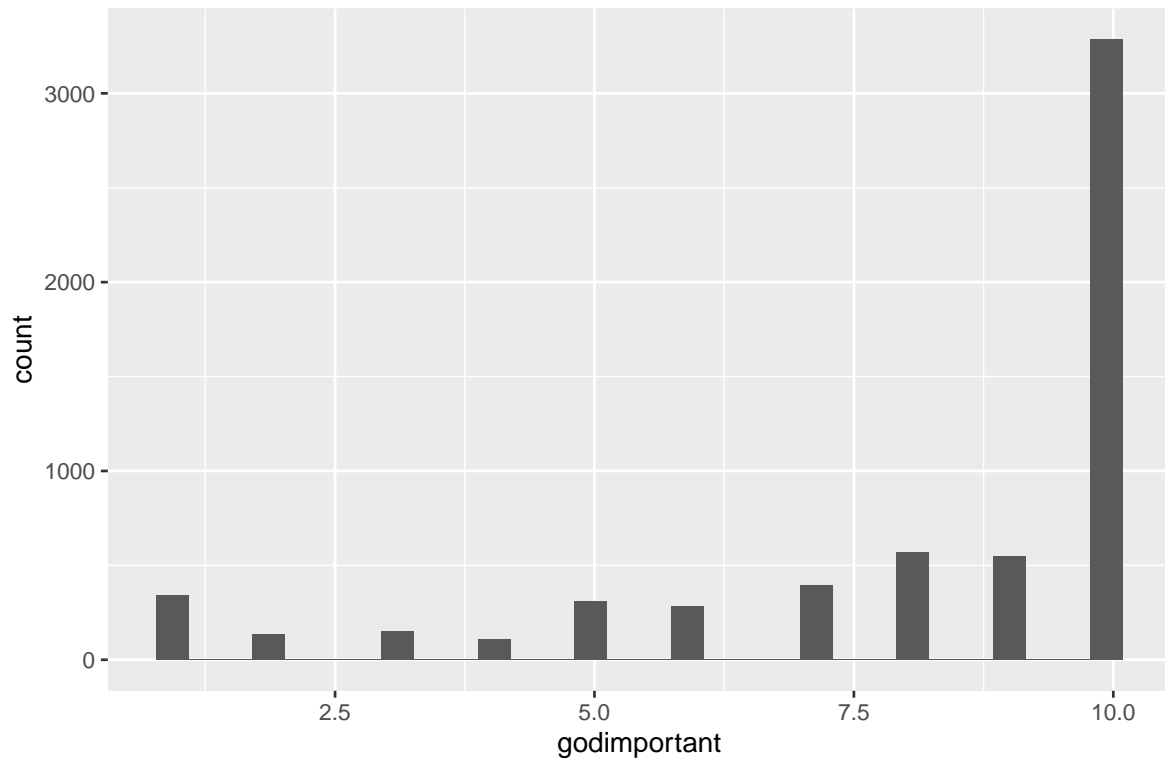
25

Histogram of respect for authority
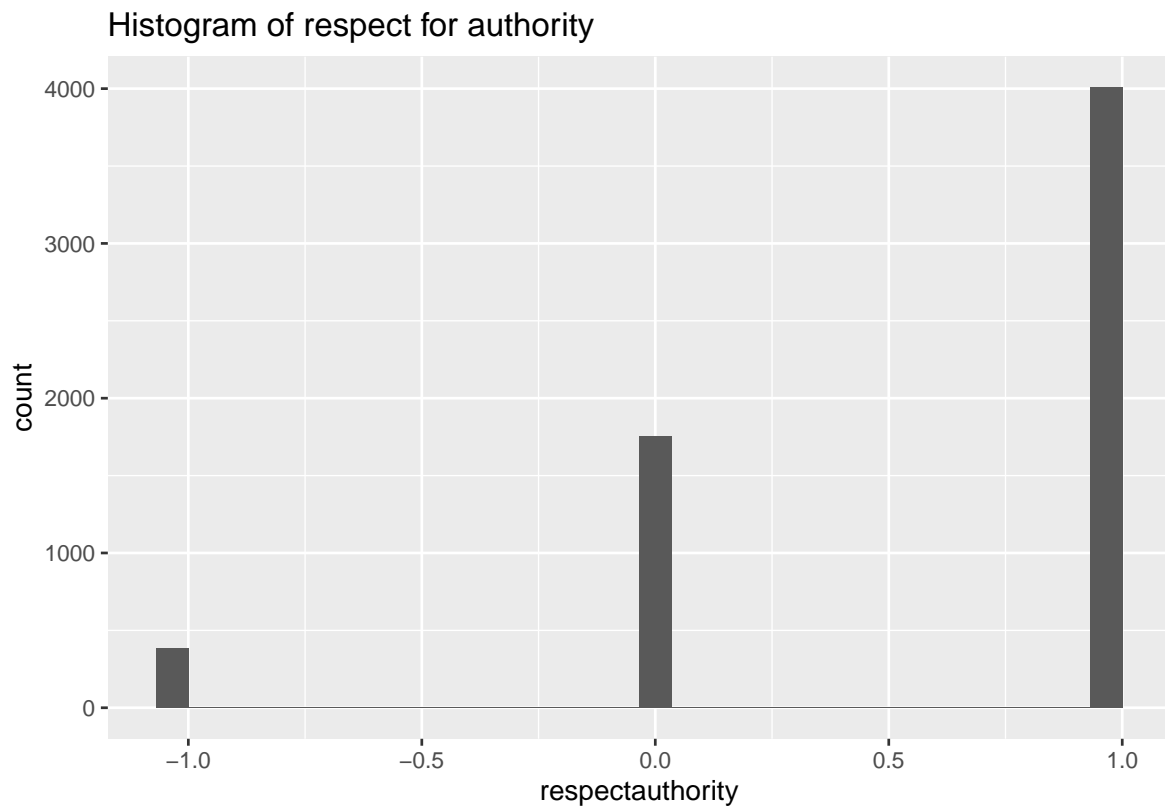

```
ggplot(abortion_data, aes(x = respectauthority, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
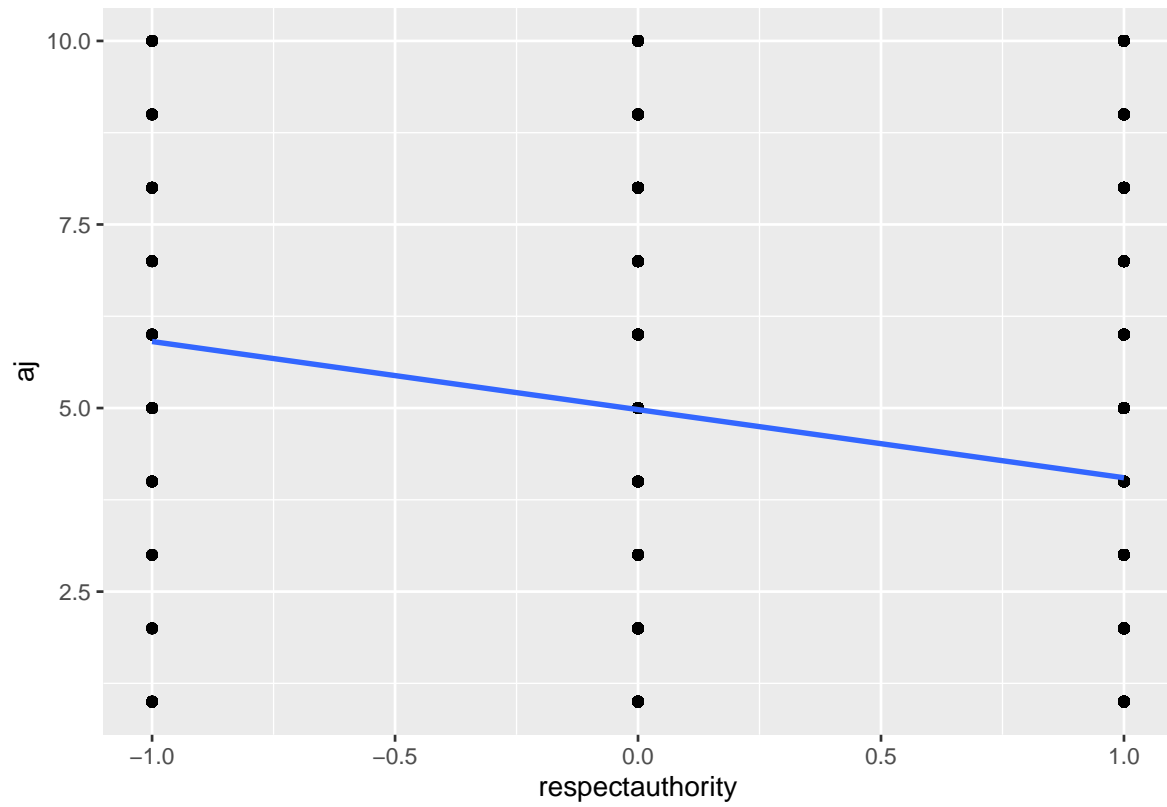
```
# National Pride
ggplot(abortion_data, aes(x = nationalpride)) +
  geom_histogram() +
  labs(title = "Histogram of National pride")
```

Histogram of National pride
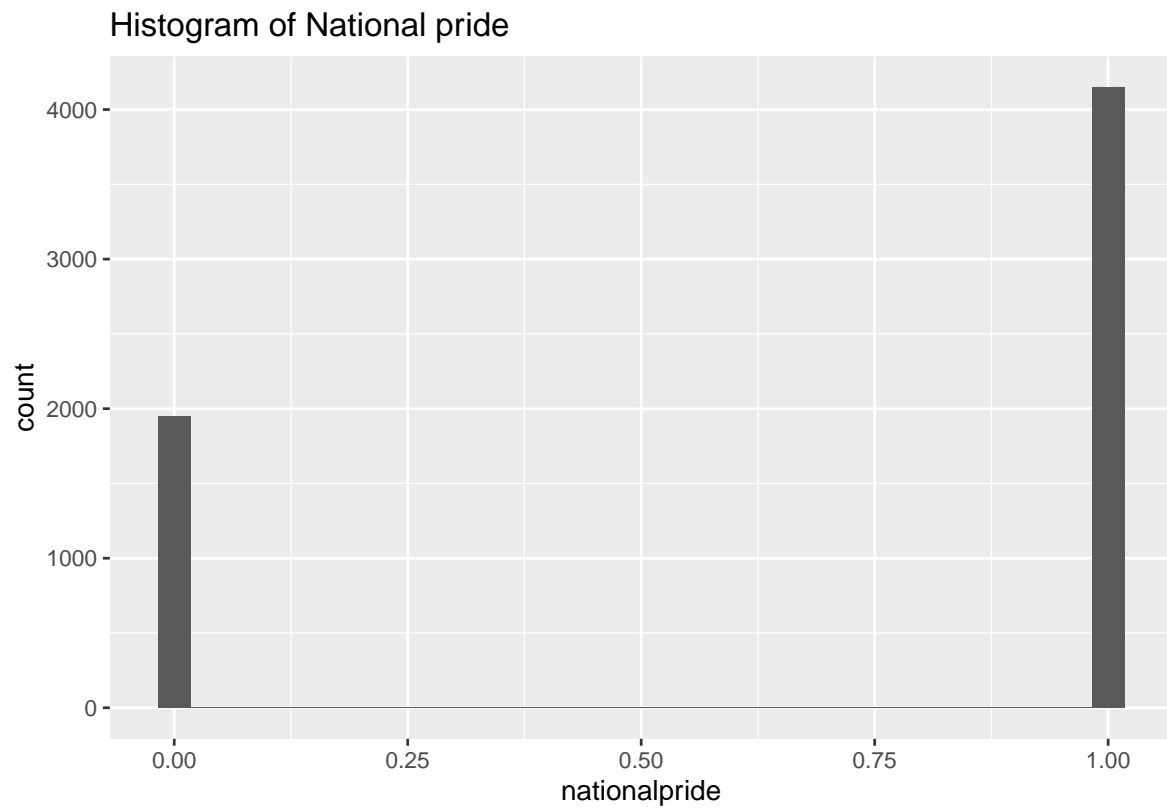

```
ggplot(abortion_data, aes(x = nationalpride, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```
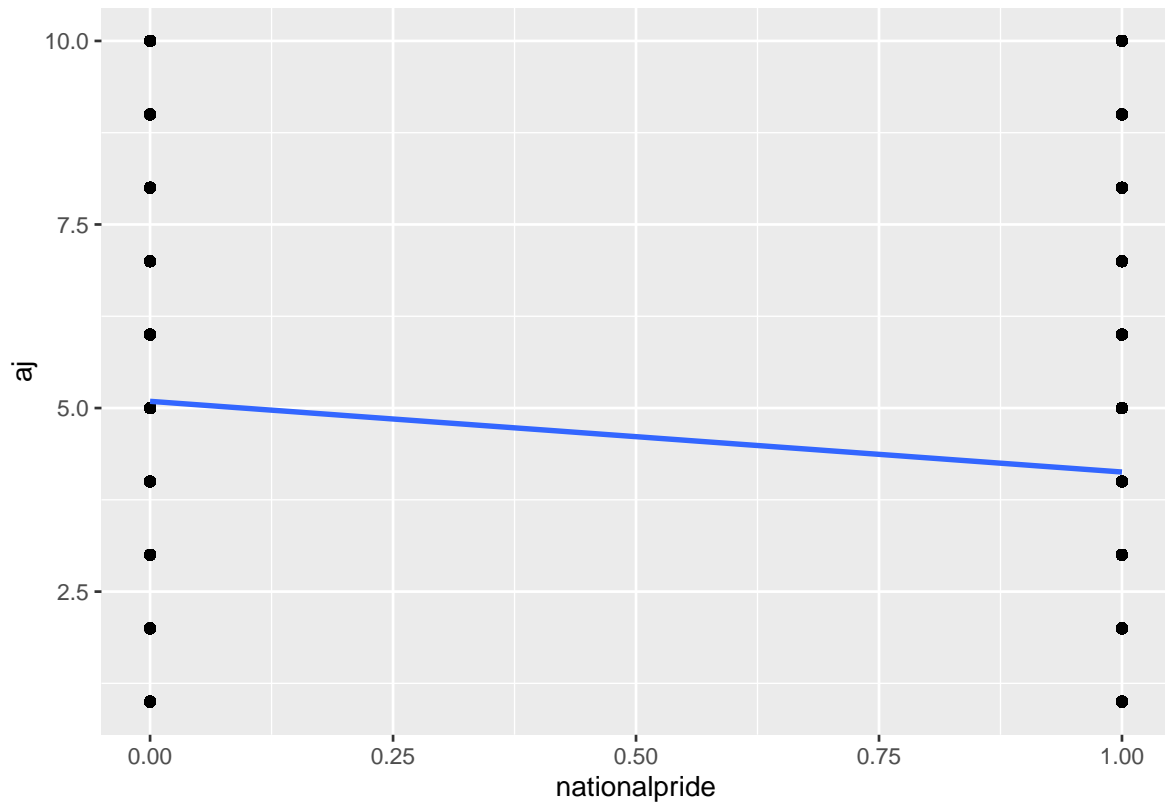
We now test if any of the predictors are strongly correlated with each other.

```
#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)
    )
```

	year	age	college	female	unemployed	ideology	fin satis.	
year		Corr: 0.059***	Corr: 0.060***	Corr: −0.012	Corr: 0.013	Corr: −0.001	Corr: −0.056***	year
age			Corr: −0.002	Corr: −0.002	Corr: −0.093***	Corr: 0.075***	Corr: 0.193***	age
college				Corr: −0.015	Corr: −0.059***	Corr: −0.055***	Corr: 0.108***	college
female					Corr: 0.031*	Corr: −0.033*	Corr: −0.014	female
unemployed						Corr: −0.018	Corr: −0.151***	unemployed
ideology							Corr: 0.145***	ideology
fin satis.								fin satis.
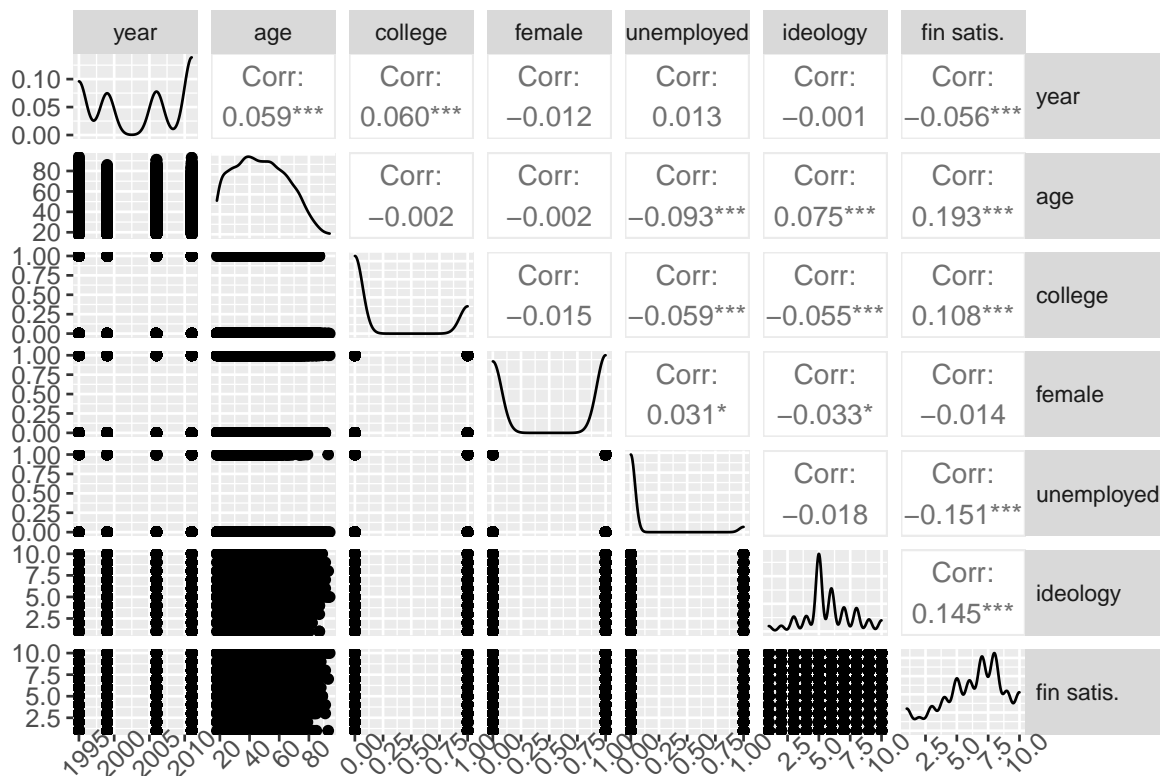
```
ggpairs(abortion_data,
  columns = c("postma4", "cai", "trustmostpeople", "godimportant",
              "respectauthority","nationalpride"),
  columnLabels = c("postmaterialism", "child autonomy", "trust", "god",
              "respectauthority","nationalpride")) +
  theme(
   axis.text.y = element_text(size = 10),
   axis.text.x = element_text(angle = 45, size = 10),
   strip.text.y = element_text(angle = 0, hjust = 0)
   )
```
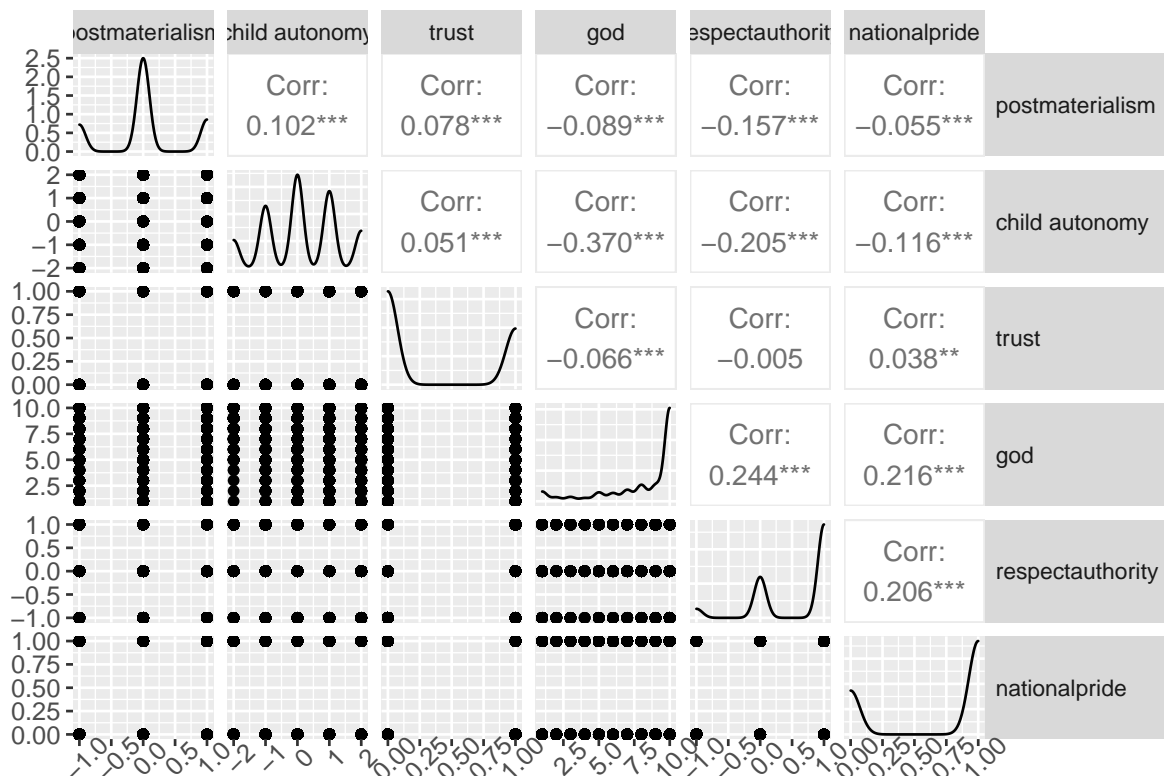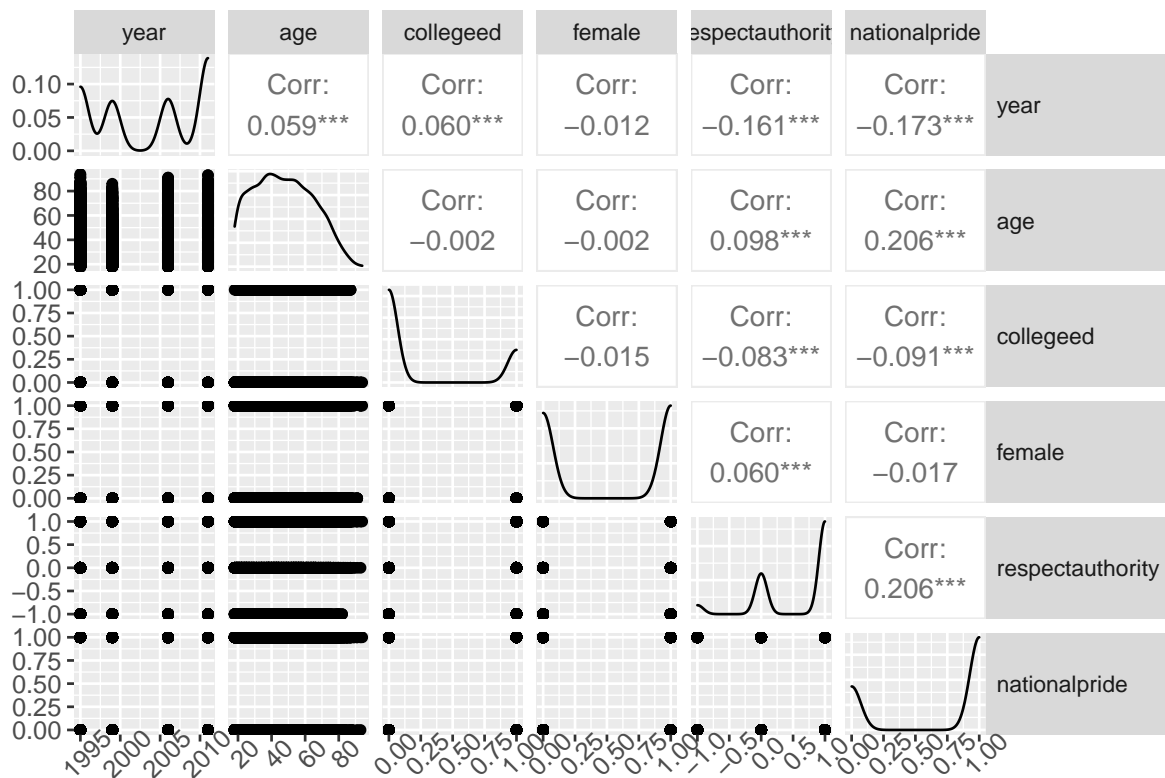
```
ggpairs(abortion_data,
  columns = c("year", "age", "collegeed", "female",
              "respectauthority","nationalpride"),
  columnLabels = c("year", "age", "collegeed", "female",
              "respectauthority","nationalpride")) +
  theme(
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(angle = 45, size = 10),
    strip.text.y = element_text(angle = 0, hjust = 0)
    )
```
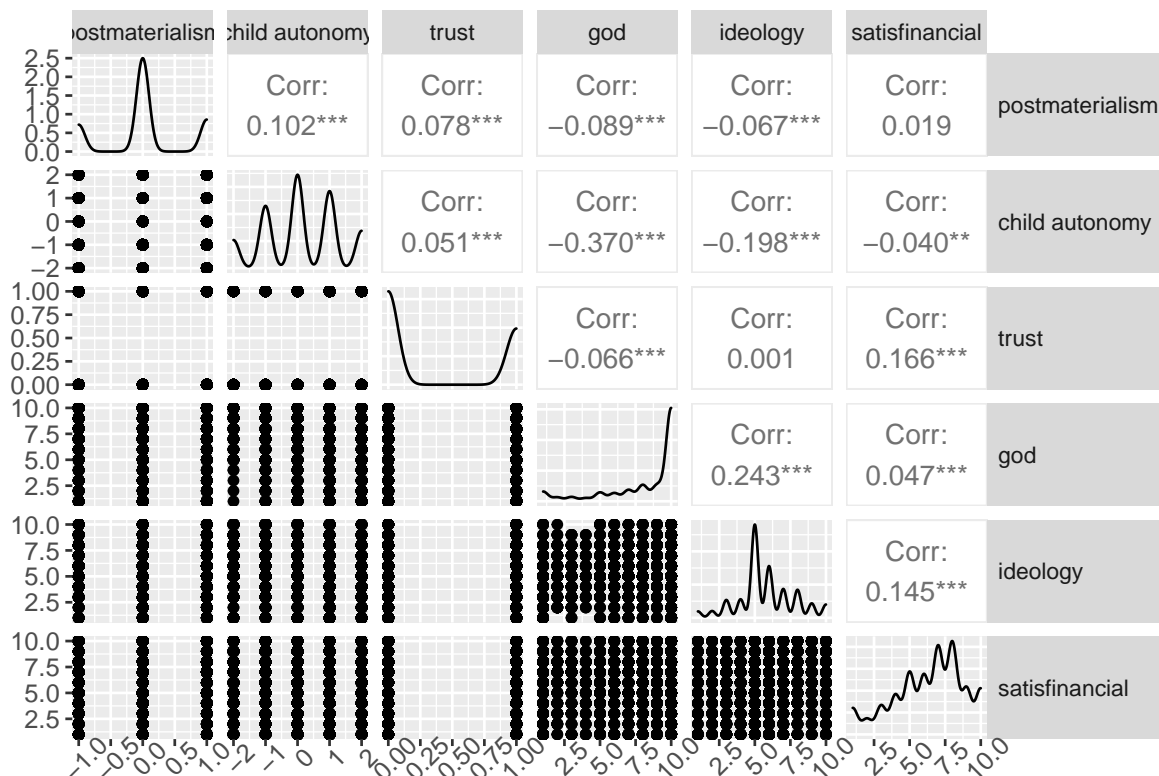
```
ggpairs(abortion_data,
  columns = c("postma4", "cai", "trustmostpeople", "godimportant",
              "ideology", "satisfinancial"),
  columnLabels = c("postmaterialism", "child autonomy", "trust", "god",
              "ideology", "satisfinancial")) +
   theme(
    axis.text.y = element_text(size = 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.

We now explore 3 different MLR models, each with 6 predictor variables chosen from the above.

```
# predictors: year, age, female, ideology, cai, respectauthority
model1_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(aj ~ year + age + female + ideology + cai + respectauthority, data = abortion_data)

glance(model1_fit) %>% select(r.squared, adj.r.squared, AIC, BIC)
```

```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
1     0.180         0.179 27214. 27268.
```

```
# predictors: year, age, ideology, cai, respectauthority, nationalpride
model2_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(aj ~ year + age + ideology + cai + respectauthority + nationalpride, data = abortion_da

glance(model2_fit) %>% select(r.squared, adj.r.squared, AIC, BIC)
```

```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
1     0.182         0.181 26999. 27052.
```

```
# predictors: year, age, female, ideology, cai, nationalpride
model3_fit <- linear_reg() %>%
  set_engine("lm") %>%
  fit(aj ~ year + age + female + ideology + cai + nationalpride, data = abortion_data)

glance(model3_fit) %>% select(r.squared, adj.r.squared, AIC, BIC)
```

```
# A tibble: 1 x 4
  r.squared adj.r.squared    AIC    BIC
      <dbl>         <dbl>  <dbl>  <dbl>
1     0.178         0.177 27152. 27205.
```

Comparing the adjusted R-squared, AIC and BIC values, it is model 2 that has the highest
adjusted R-squared and the lowest AIC and BIC of the 3 models. **Thus, we will use
model 2 with the following predictors: year, age, ideology, cai, respectauthority,
nationalpride.**

Data dictionary

The data dictionary can be found here.