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

[Christina Wang]

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

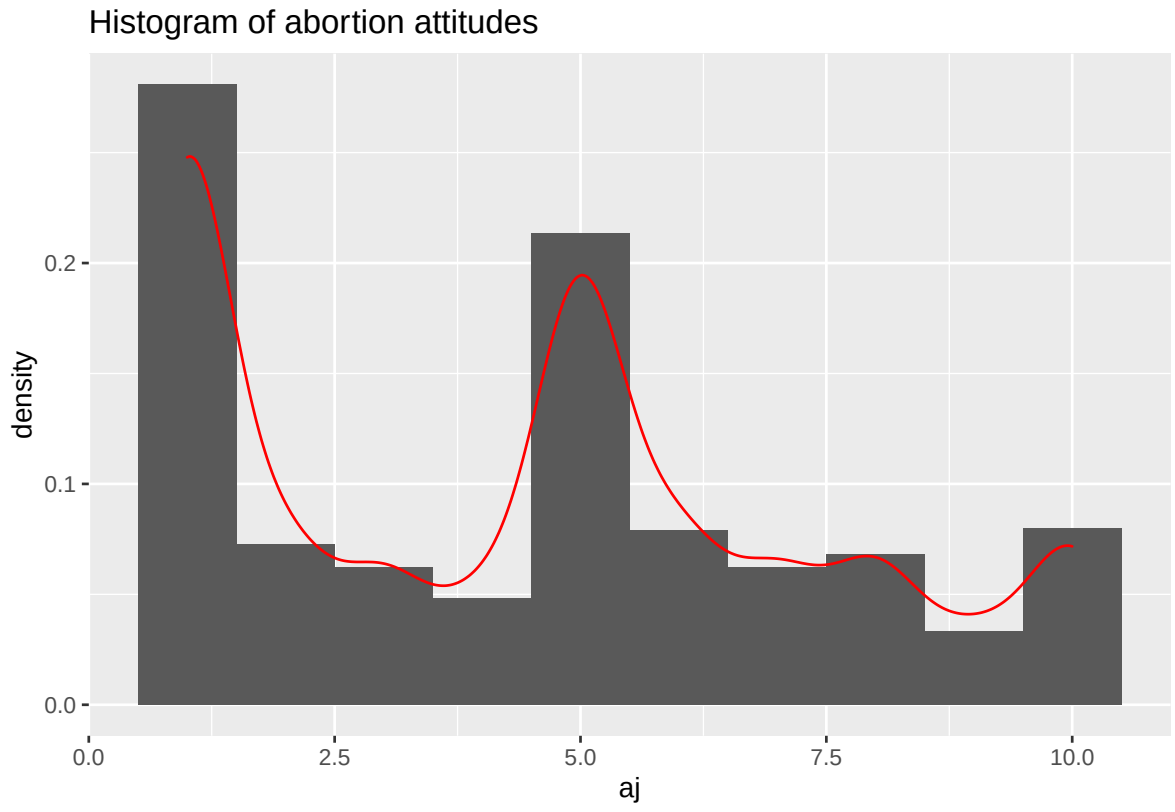
remove the observations in generational waves before 1995 so that our data includes all the survey questions.

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")
```

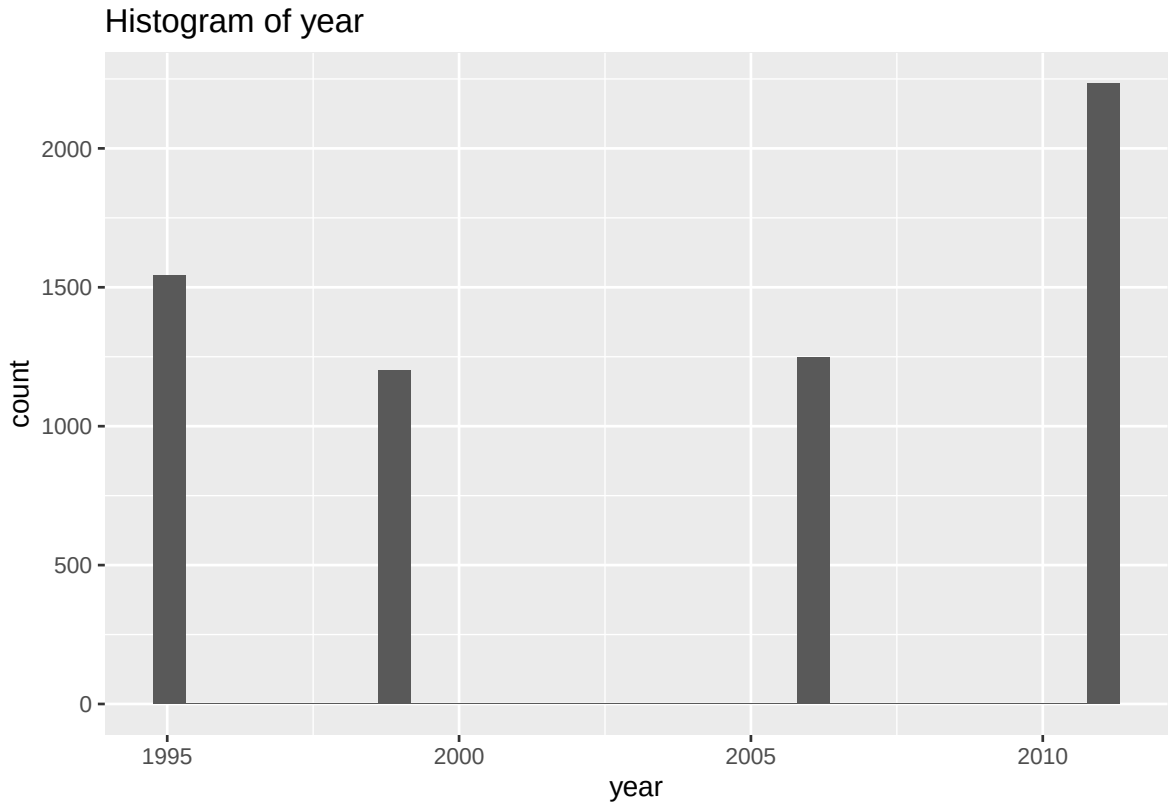


```
summary(abortion_data$aj)
```

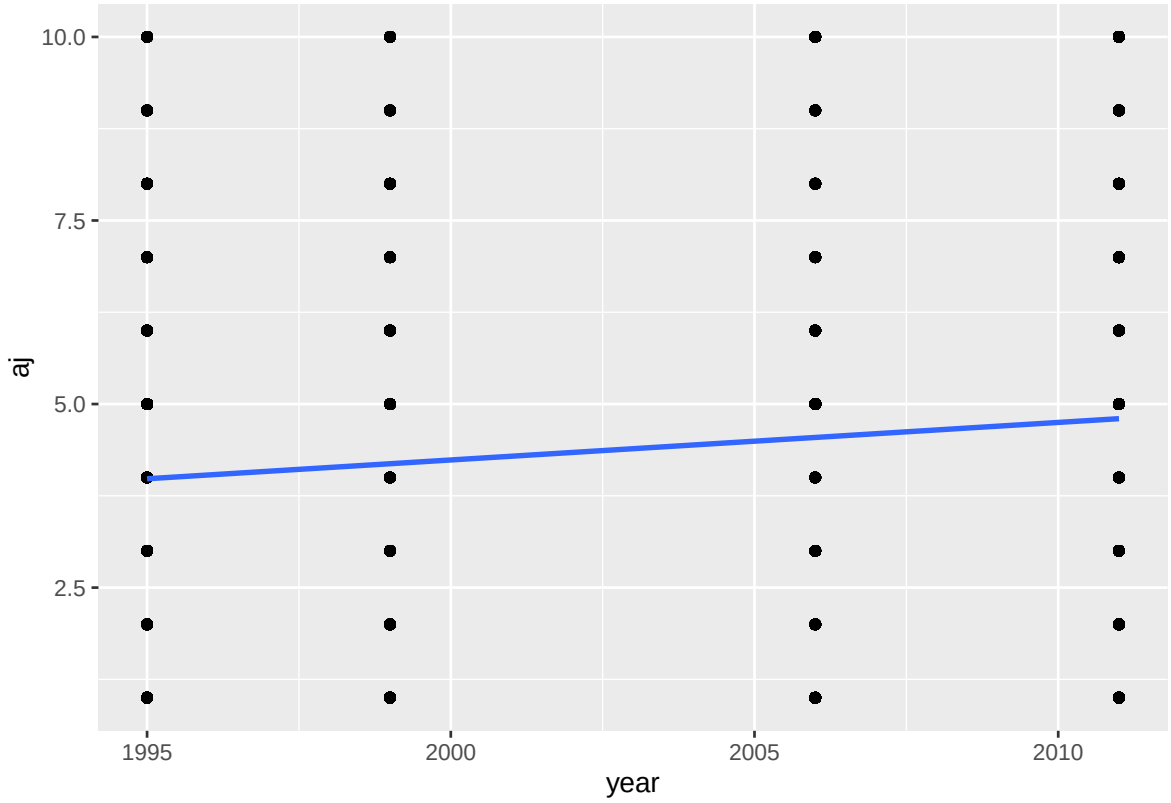
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.000   5.000   4.428   6.000  10.000     214
```

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.

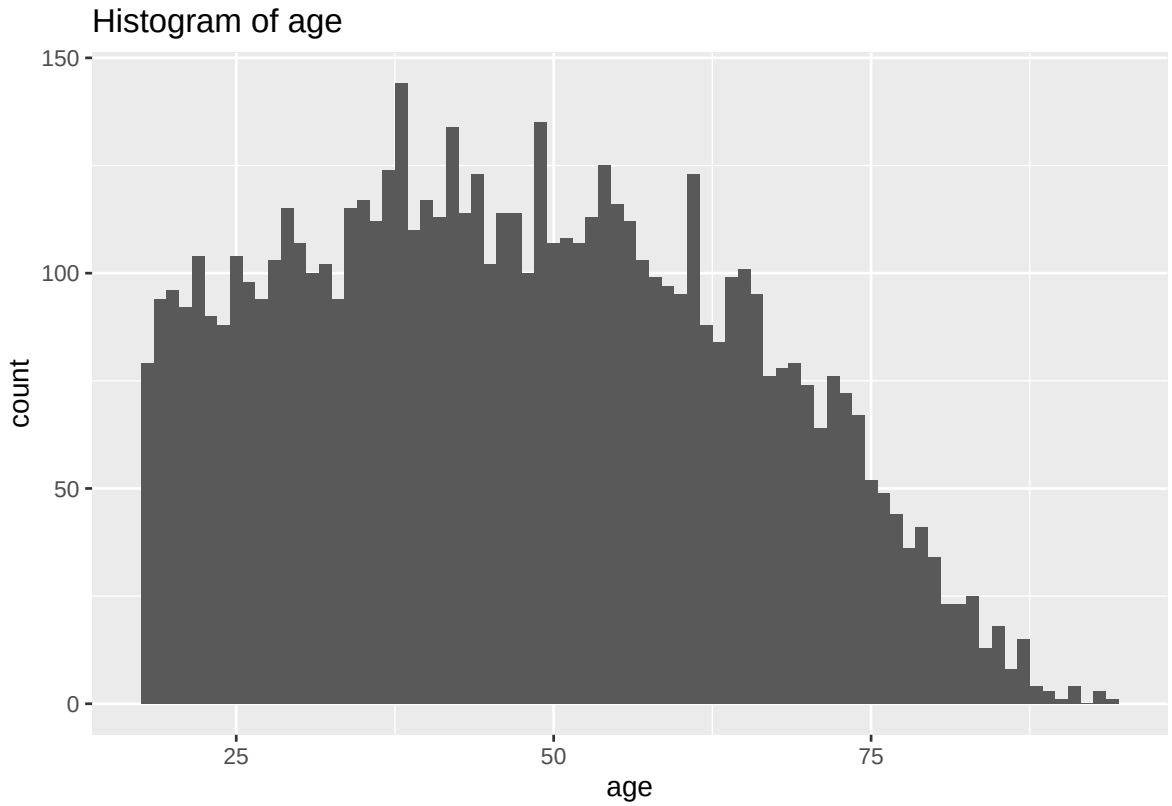
```
#Year of survey
ggplot(abortion_data, aes(x = year)) +
  geom_histogram() +
  labs(title = "Histogram of year")
```



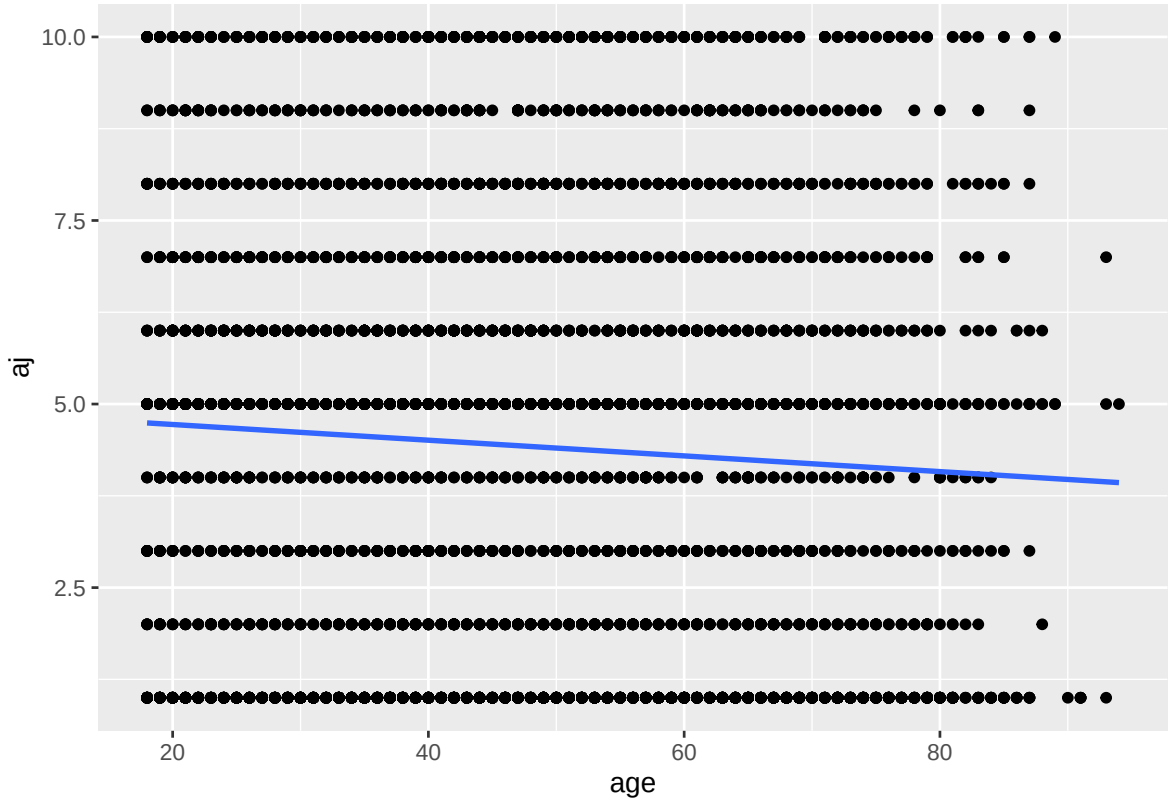

```
ggplot(abortion_data, aes(x = year, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



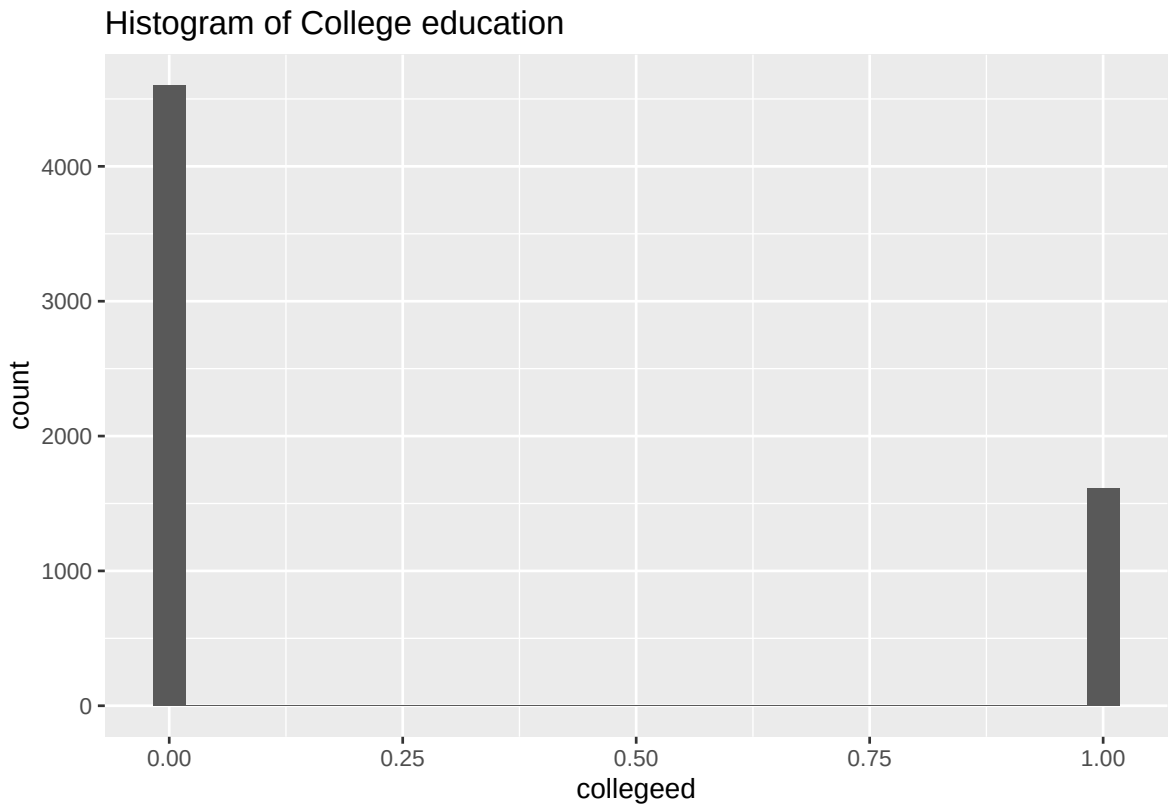

```
#Age
ggplot(abortion_data, aes(x = age)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Histogram of age")
```



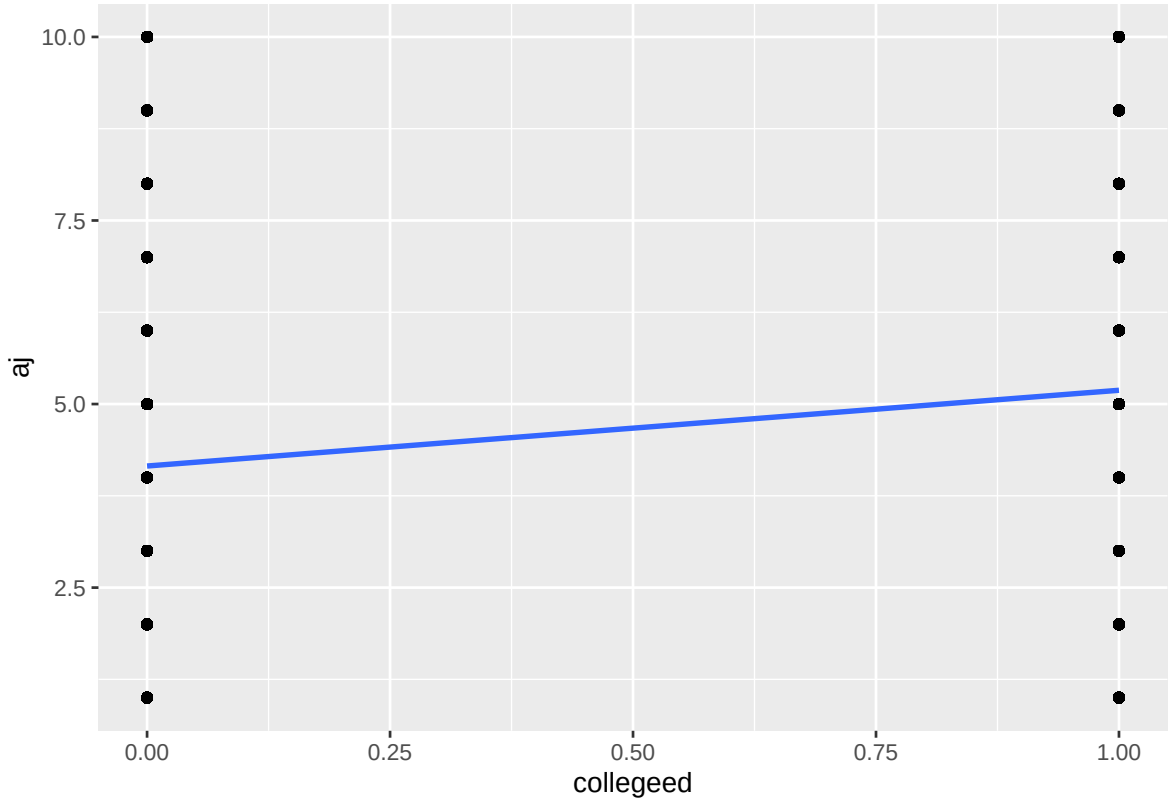

```
ggplot(abortion_data, aes(x = age, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



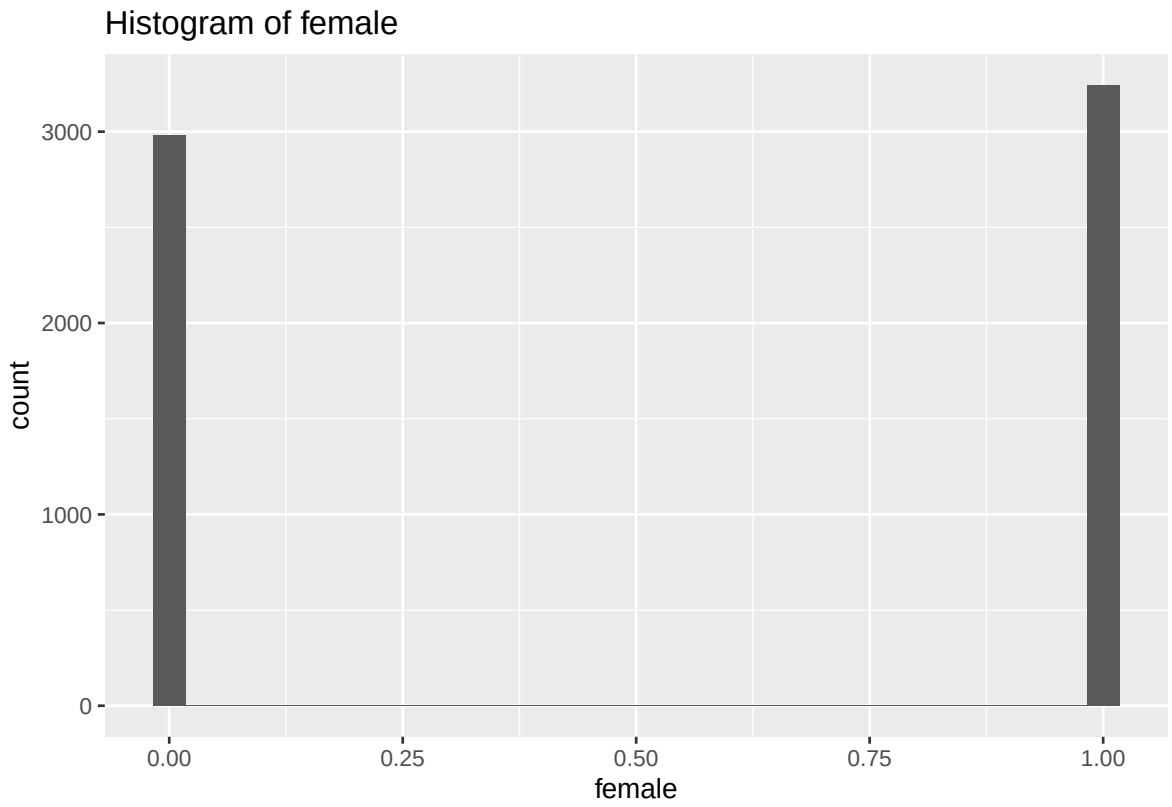

```
# College education
ggplot(abortion_data, aes(x = collegeed)) +
  geom_histogram() +
  labs(title = "Histogram of College education")
```



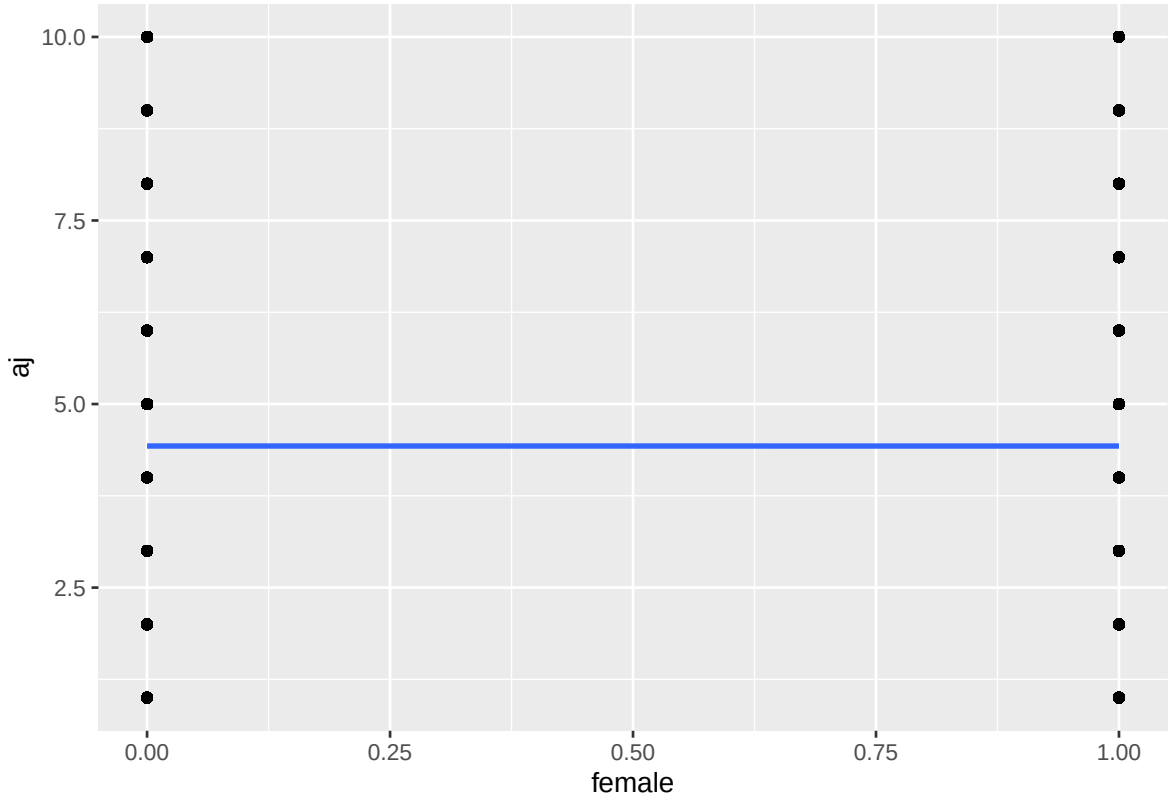

```
ggplot(abortion_data, aes(x = collegeed, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



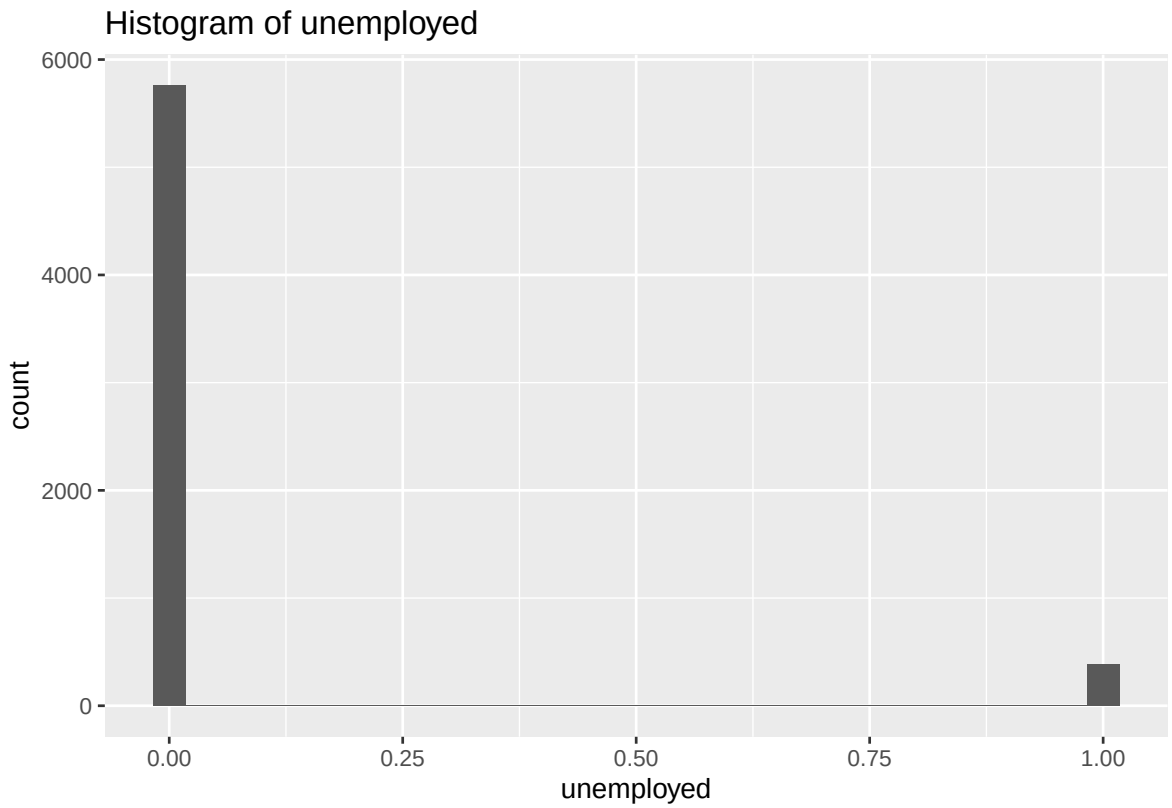

```
# Female
ggplot(abortion_data, aes(x = female)) +
  geom_histogram() +
  labs(title = "Histogram of female")
```



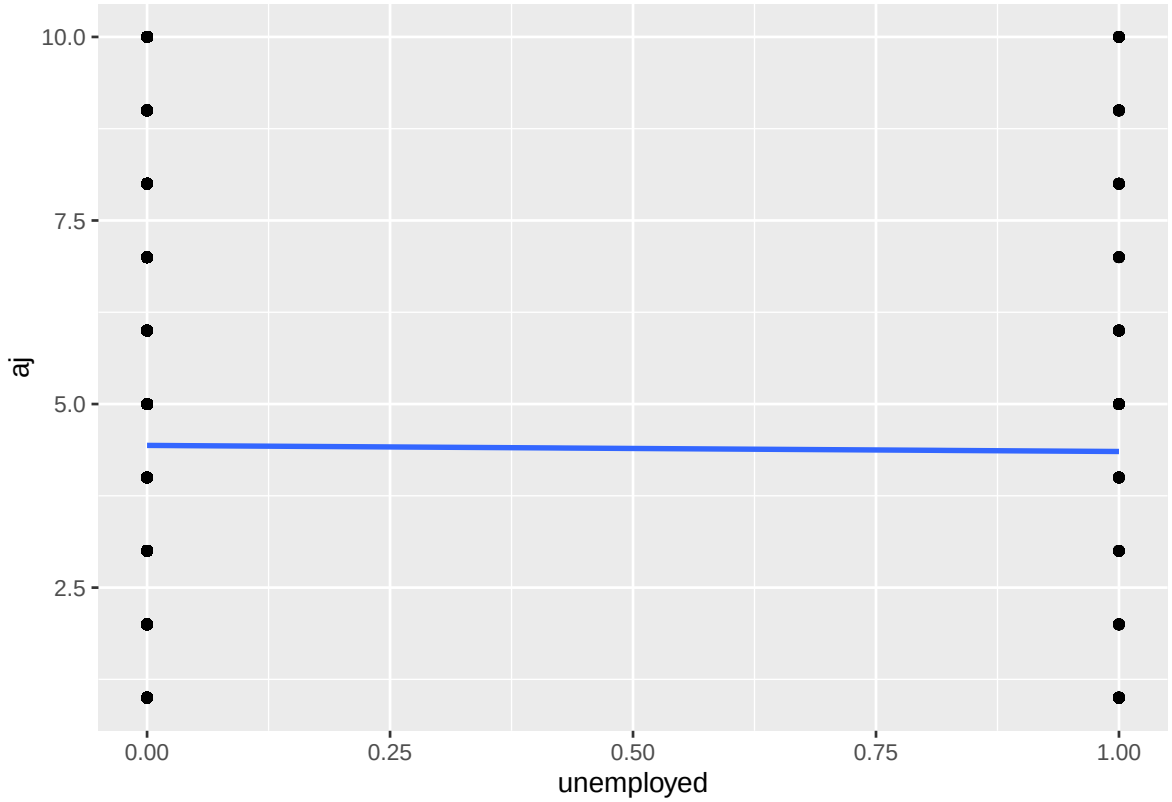

```
ggplot(abortion_data, aes(x = female, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



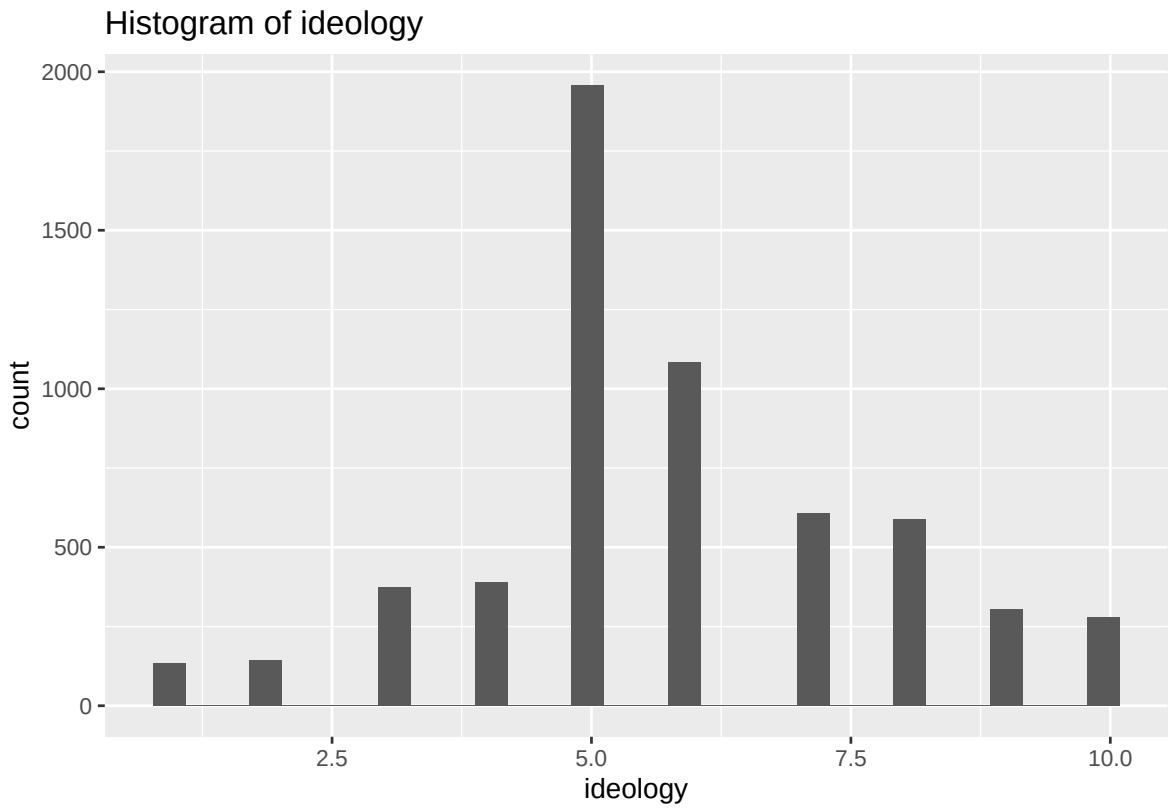

```
# Unemployed
ggplot(abortion_data, aes(x = unemployed)) +
  geom_histogram() +
  labs(title = "Histogram of unemployed")
```



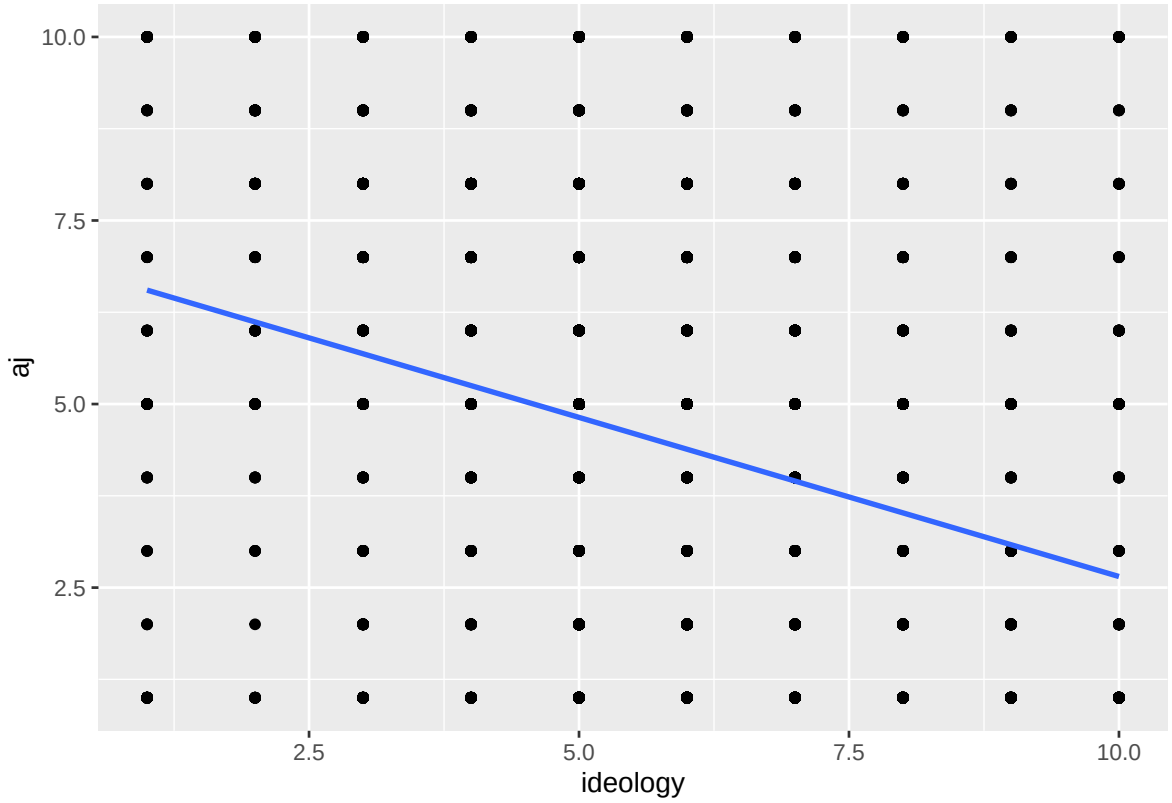

```
ggplot(abortion_data, aes(x = unemployed, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



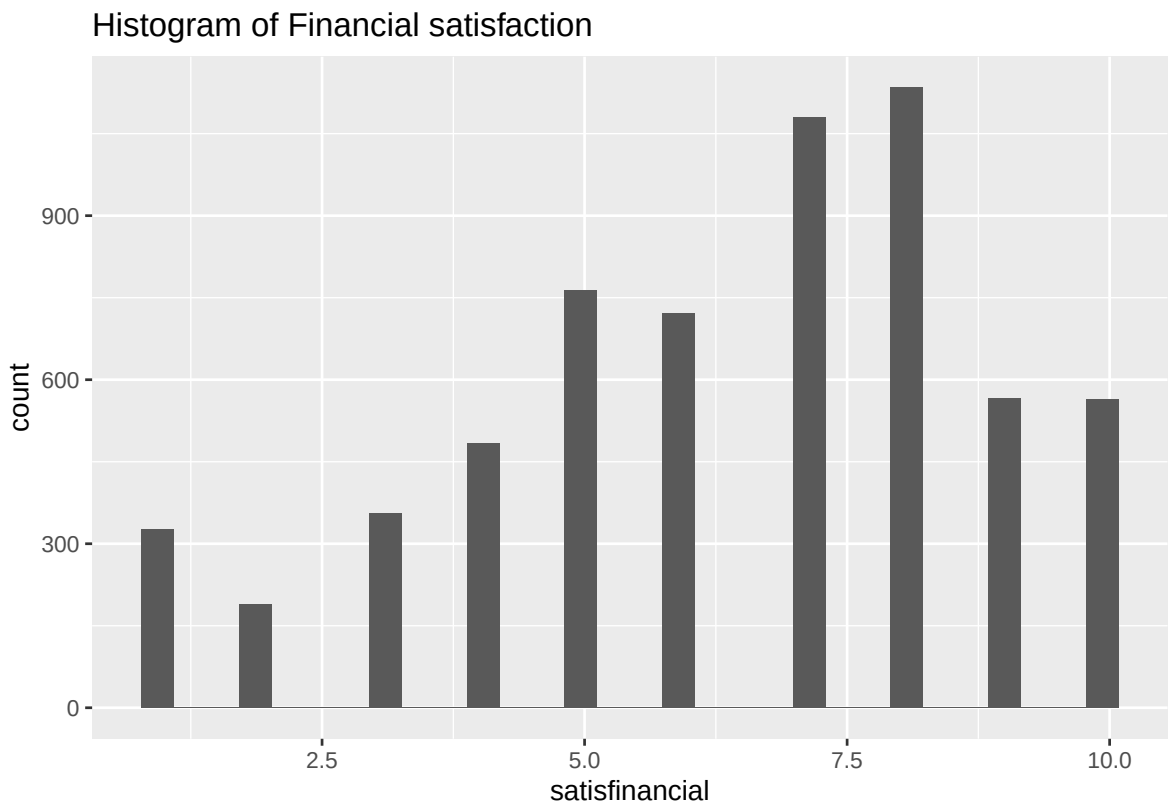

```
# ideology
ggplot(abortion_data, aes(x = ideology)) +
  geom_histogram() +
  labs(title = "Histogram of ideology")
```



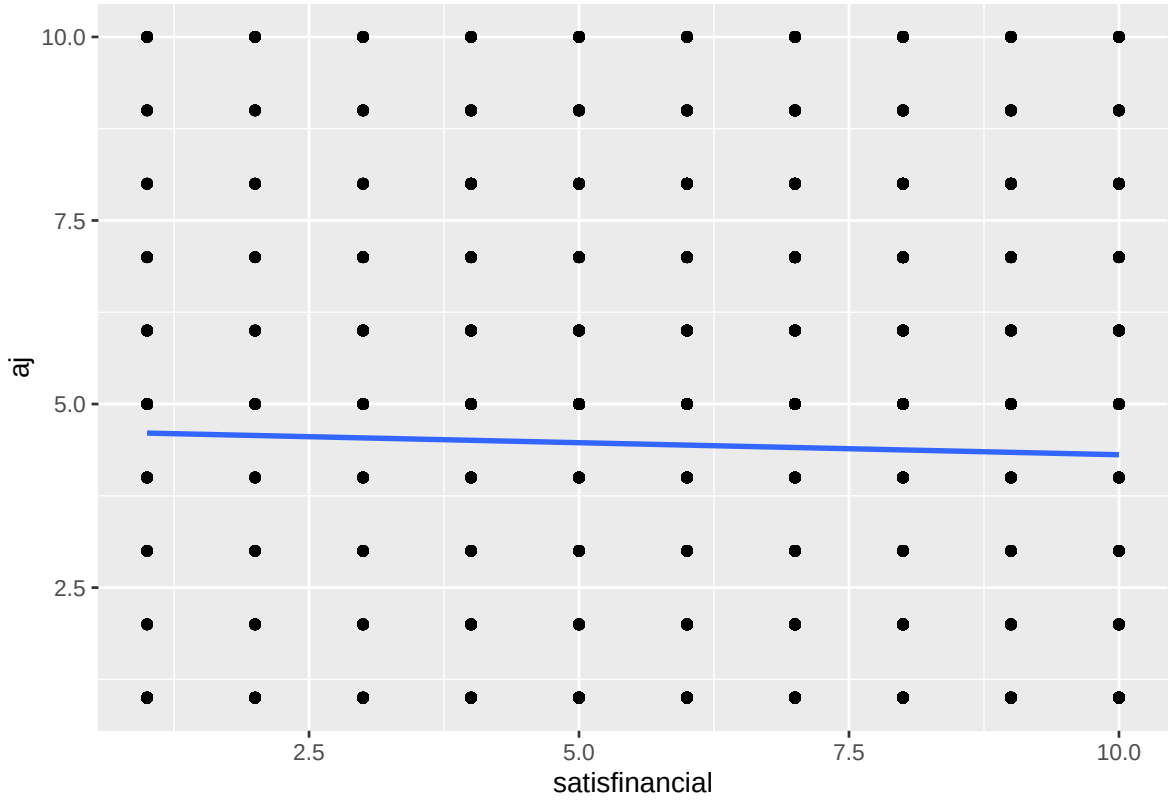

```
ggplot(abortion_data, aes(x = ideology, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



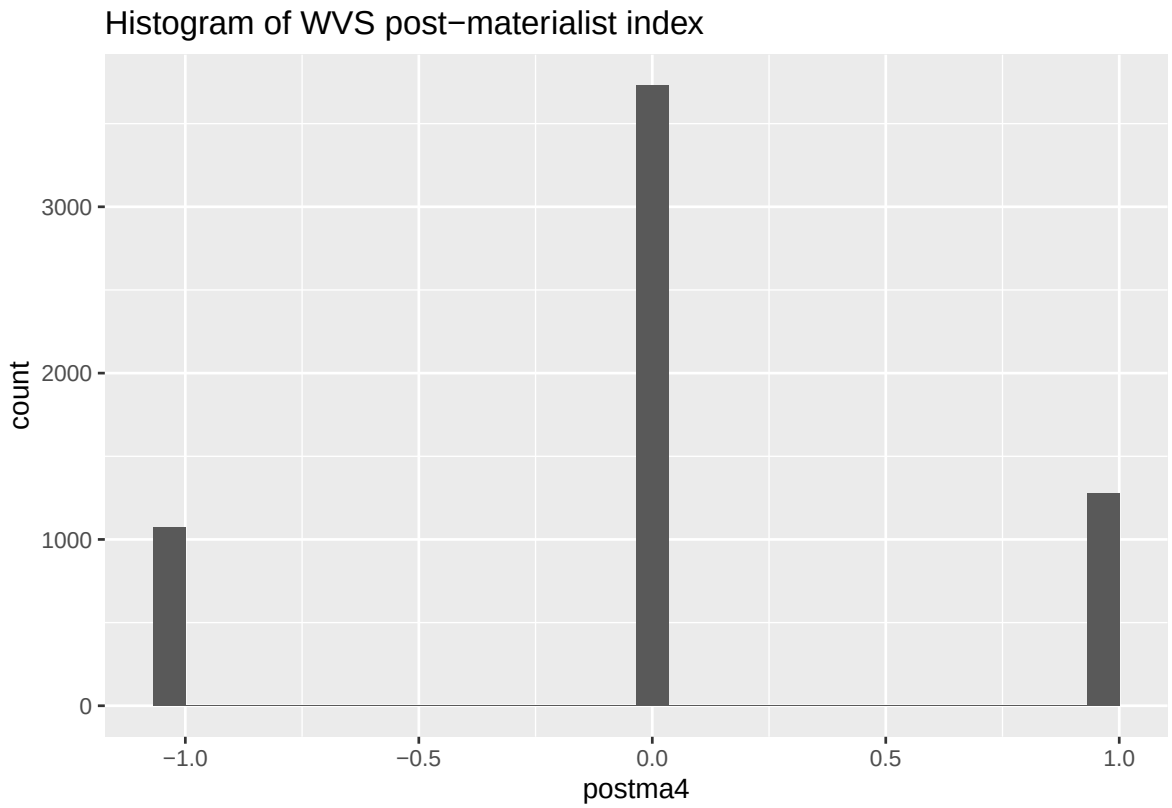

```
# Financial satisfaction
ggplot(abortion_data, aes(x = satisfinancial)) +
  geom_histogram() +
  labs(title = "Histogram of Financial satisfaction")
```



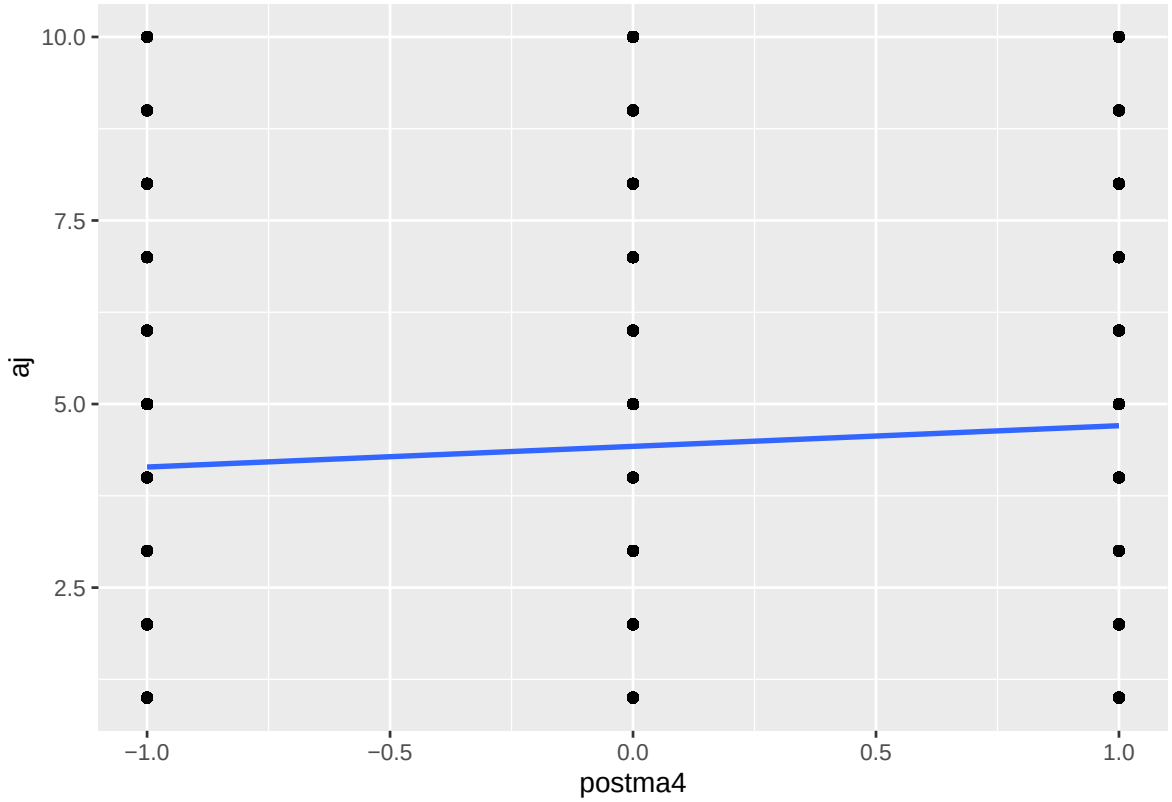

```
ggplot(abortion_data, aes(x = satisfinancial, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



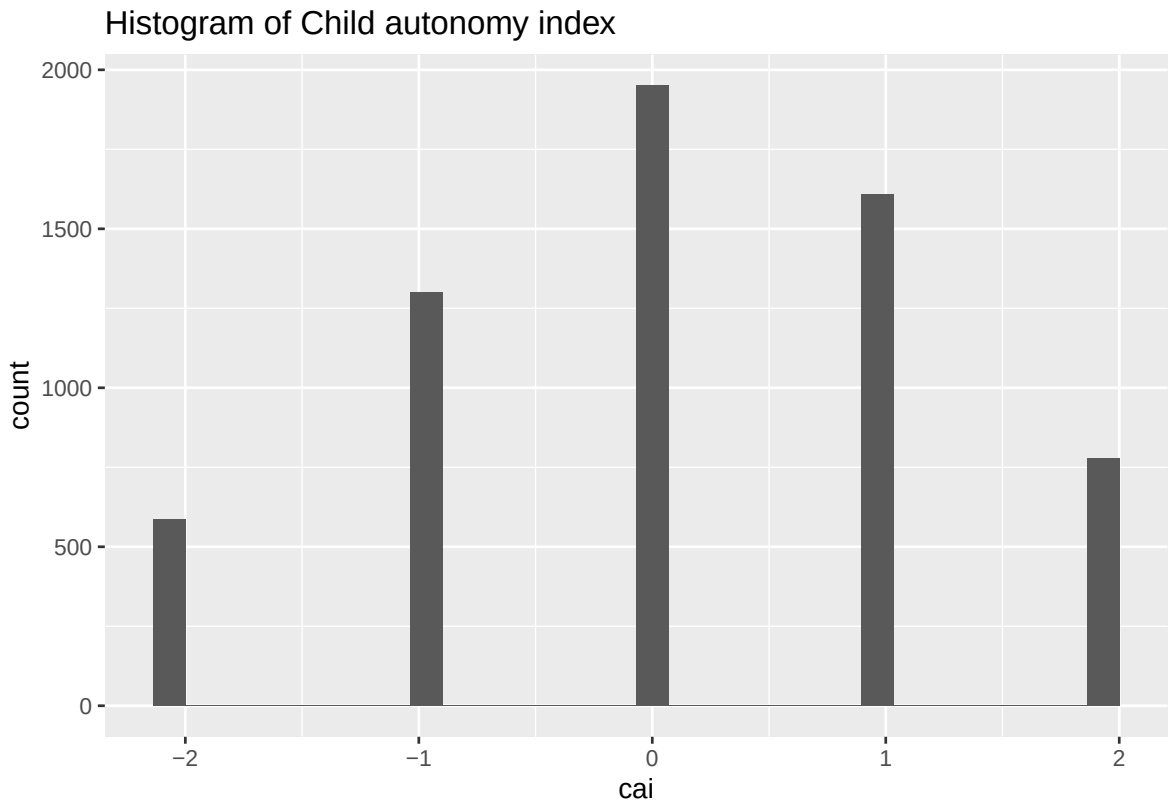

```
# WVS post-materialist index
ggplot(abortion_data, aes(x = postma4)) +
  geom_histogram() +
  labs(title = "Histogram of WVS post-materialist index")
```



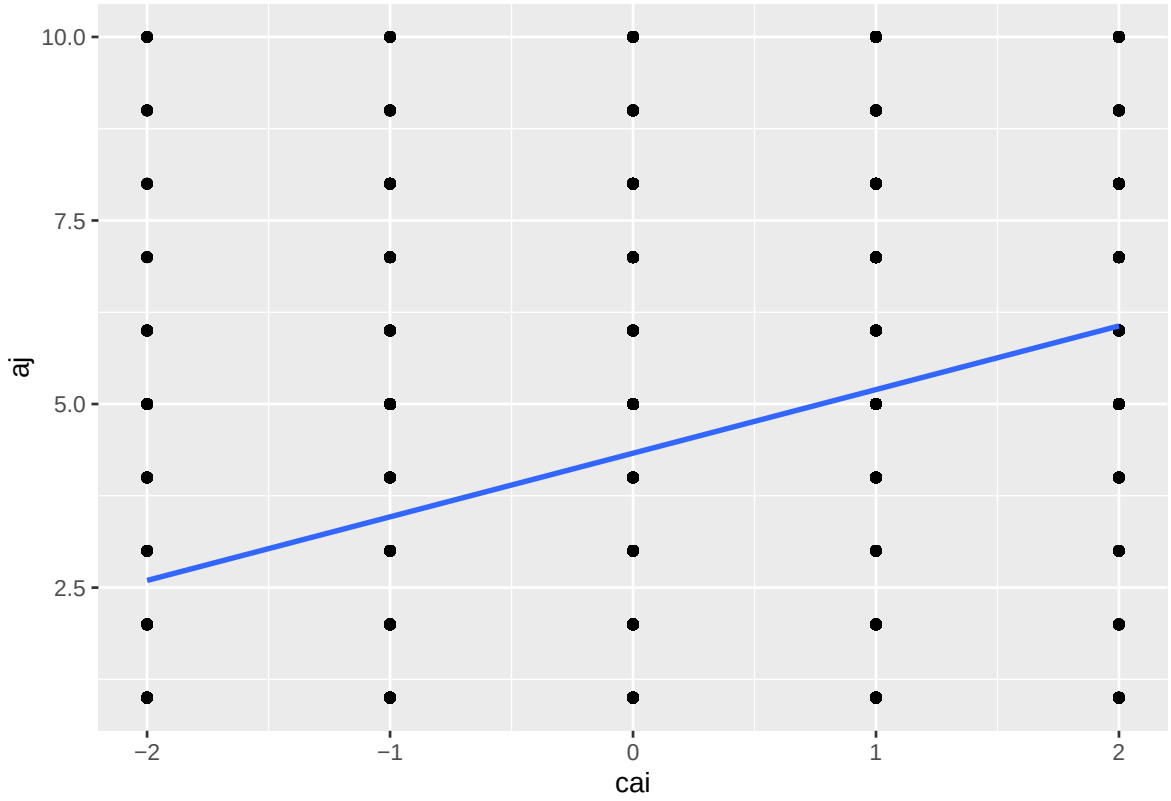

```
ggplot(abortion_data, aes(x = postma4, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



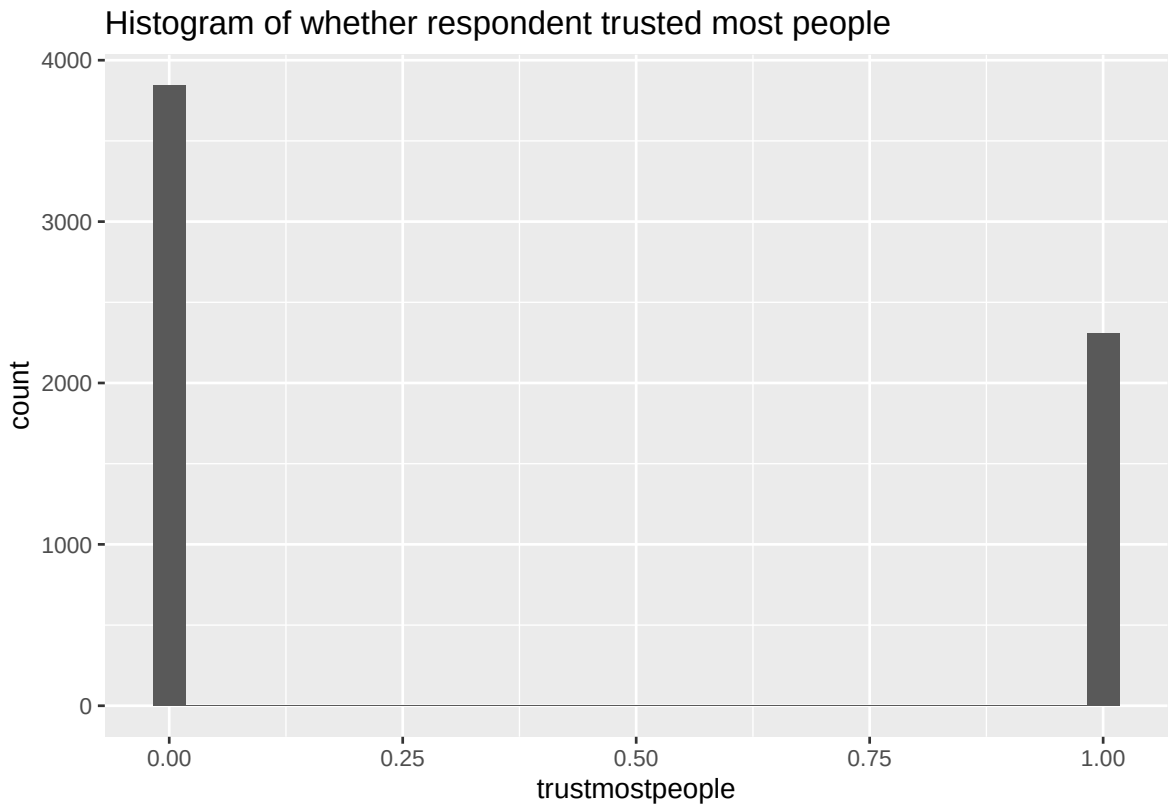

```
# Child autonomy index
ggplot(abortion_data, aes(x = cai)) +
  geom_histogram() +
  labs(title = "Histogram of Child autonomy index")
```



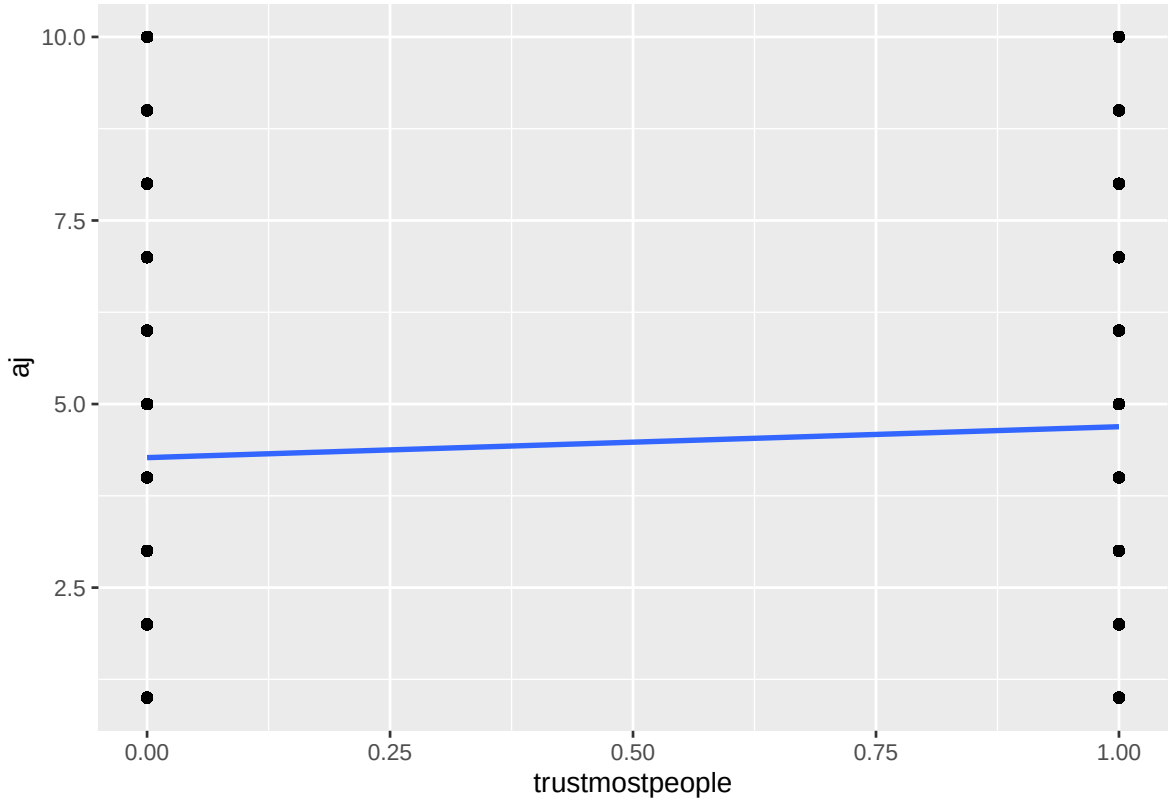

```
ggplot(abortion_data, aes(x = cai, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



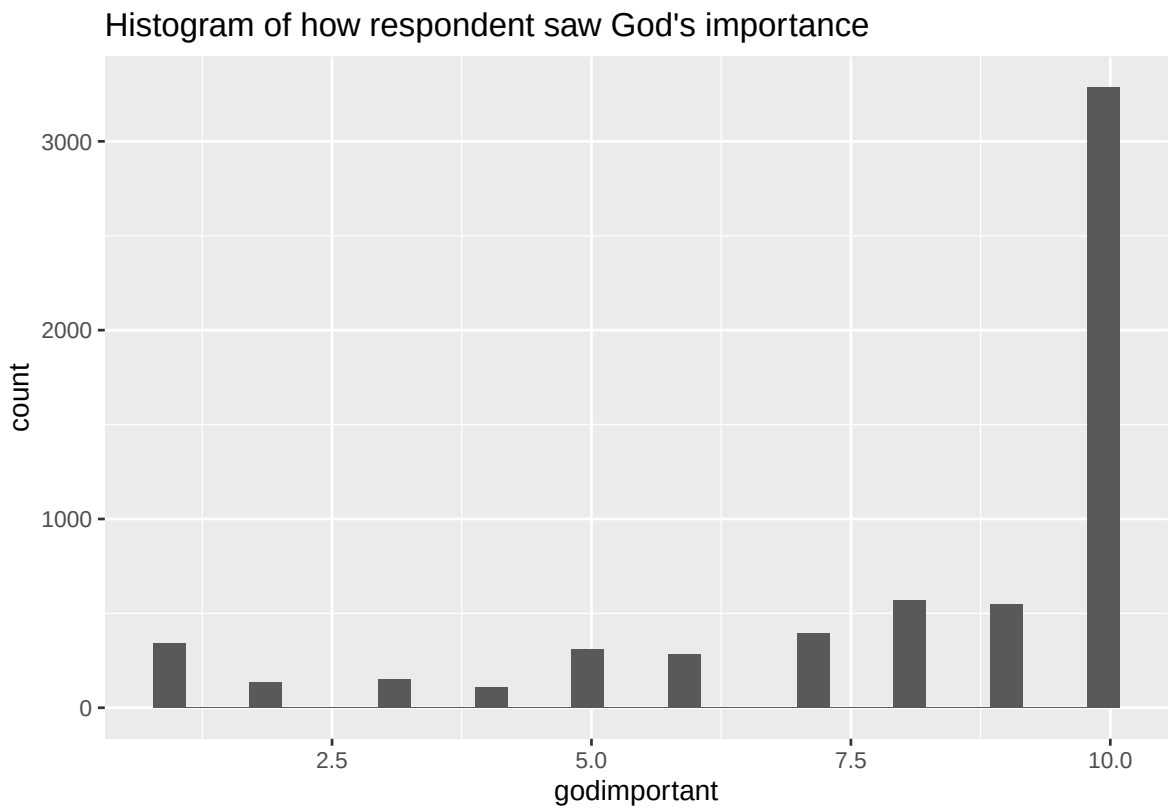

```
# Trust most people
ggplot(abortion_data, aes(x = trustmostpeople)) +
  geom_histogram() +
  labs(title = "Histogram of whether respondent trusted most people")
```



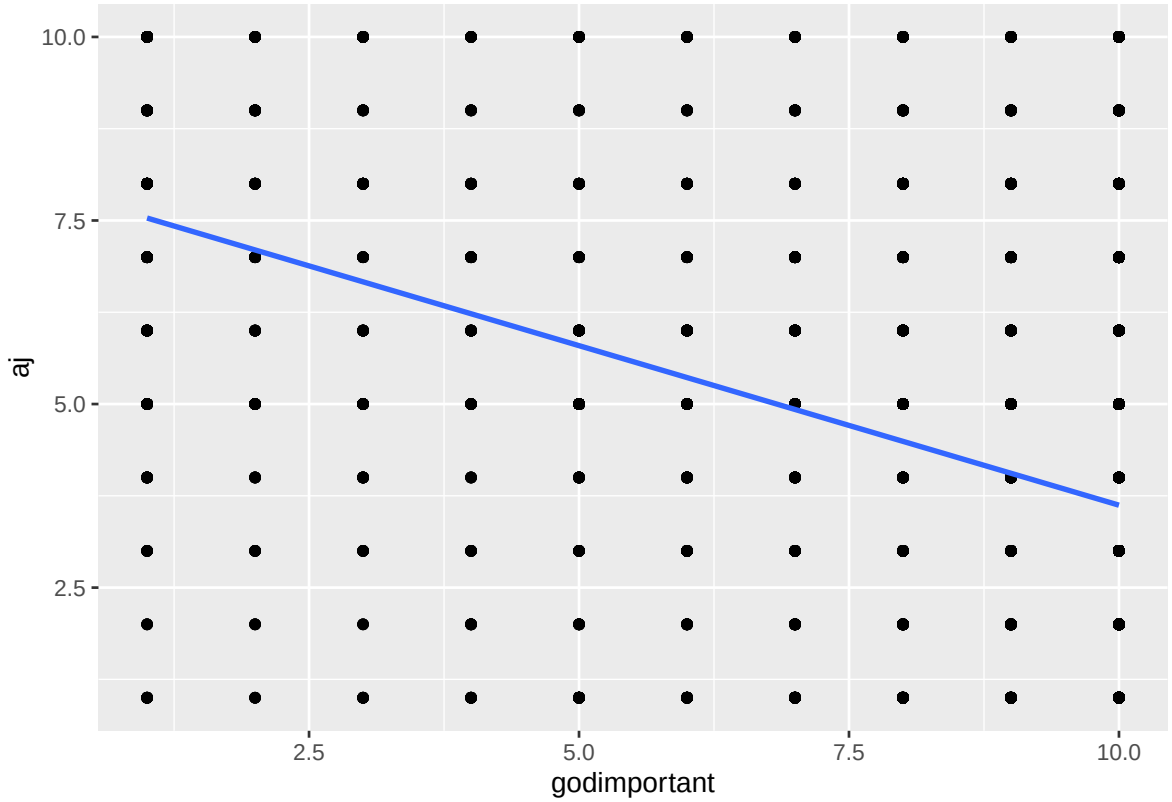

```
ggplot(abortion_data, aes(x = trustmostpeople, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



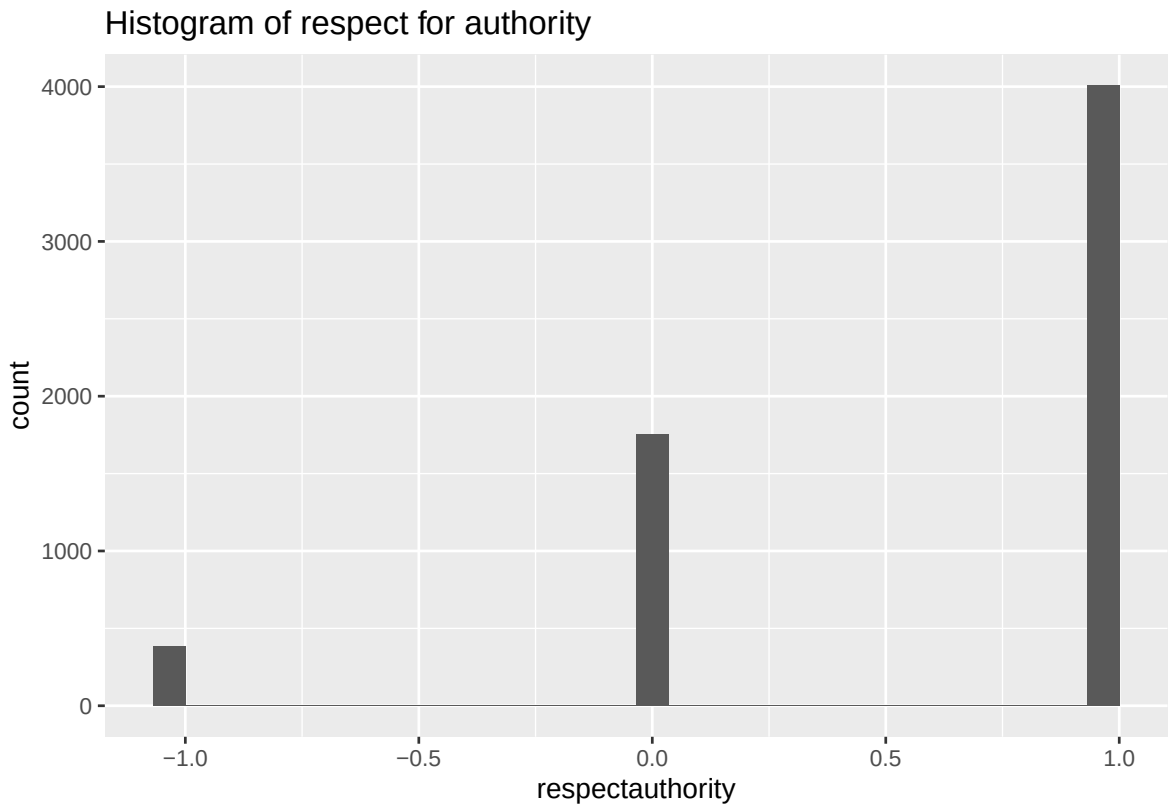

```
# Importance of God
ggplot(abortion_data, aes(x = godimportant)) +
  geom_histogram() +
  labs(title = "Histogram of how respondent saw God's importance")
```



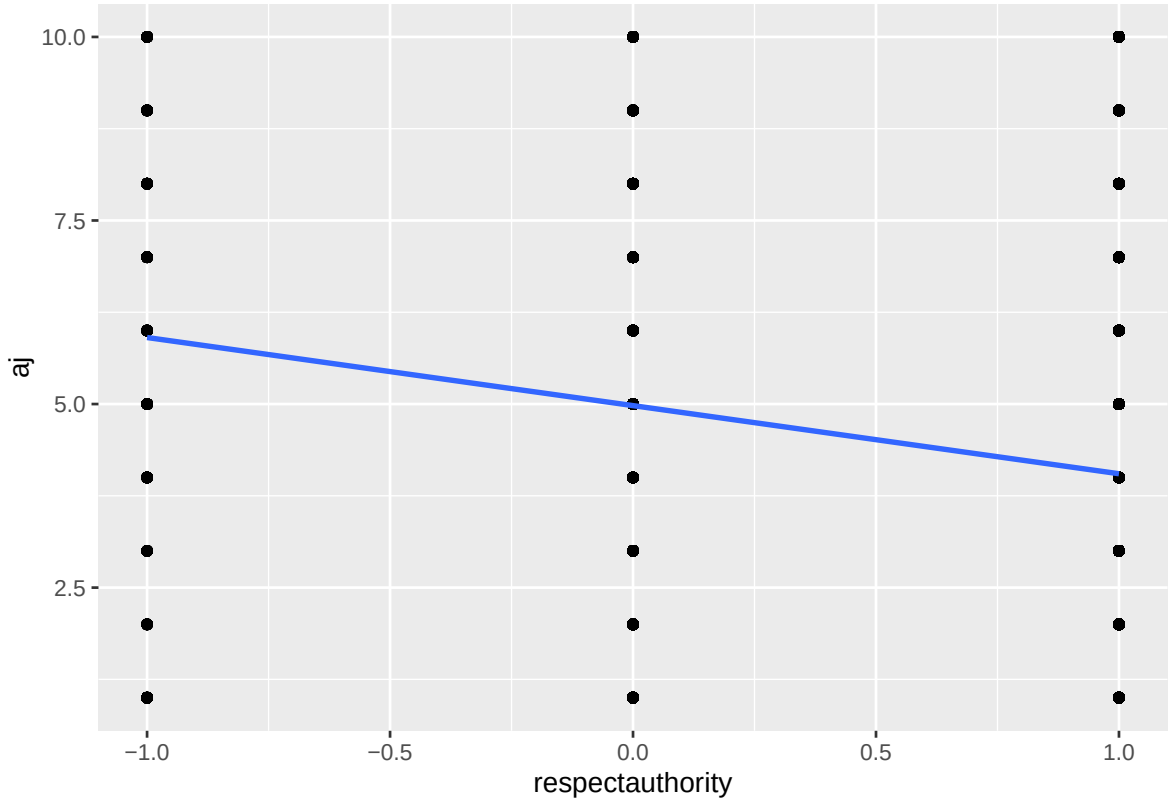

```
ggplot(abortion_data, aes(x = godimportant, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



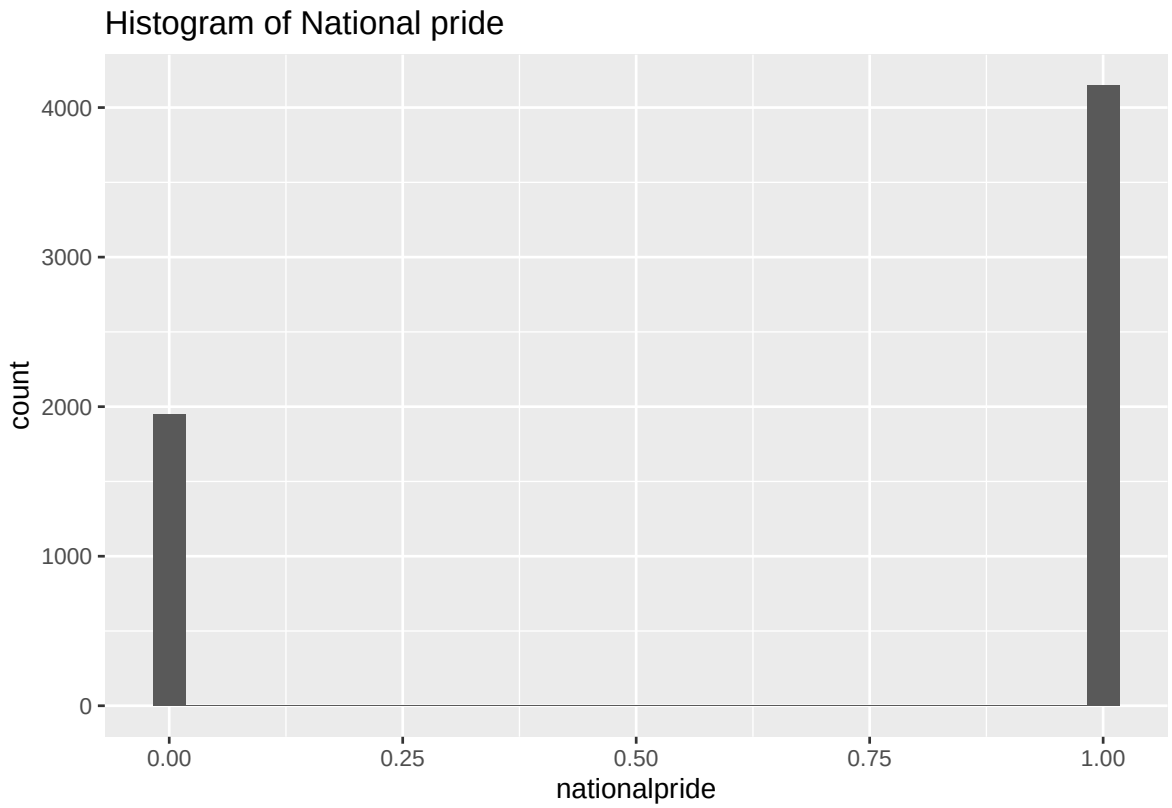

```
# Respect for authority
ggplot(abortion_data, aes(x = respectauthority)) +
  geom_histogram() +
  labs(title = "Histogram of respect for authority")
```



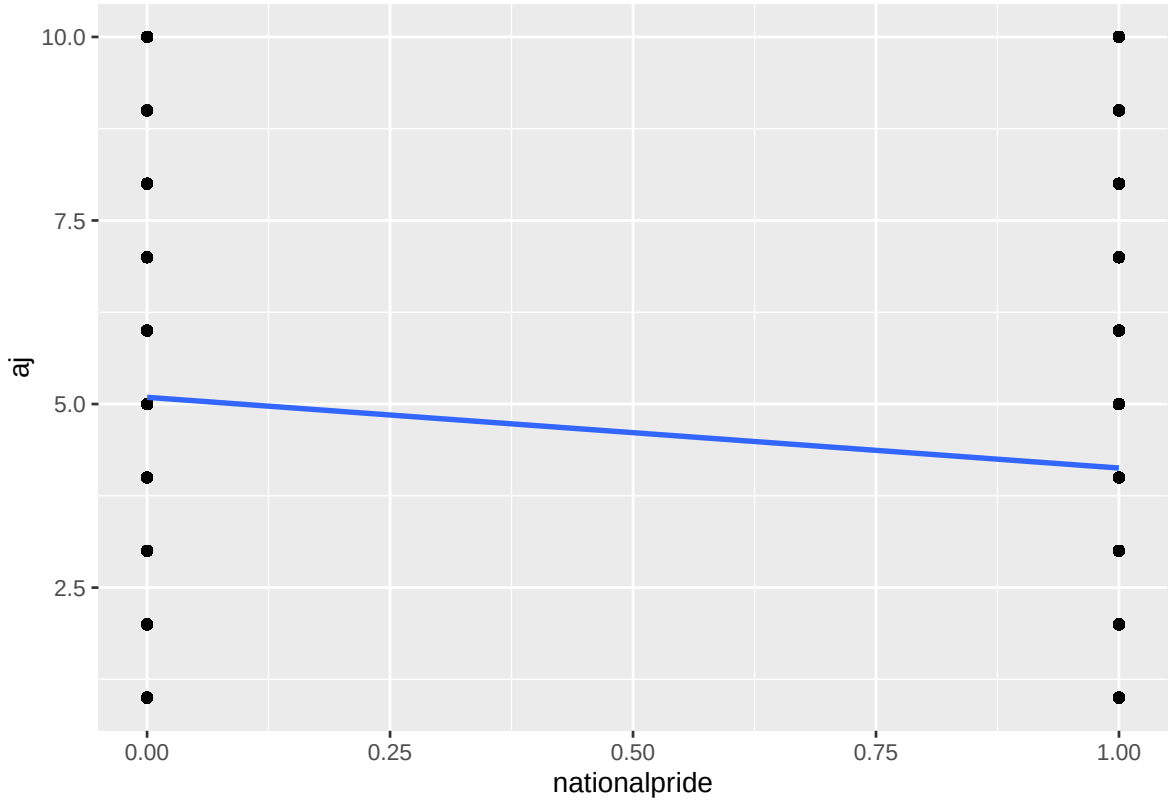

```
ggplot(abortion_data, aes(x = respectauthority, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```




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

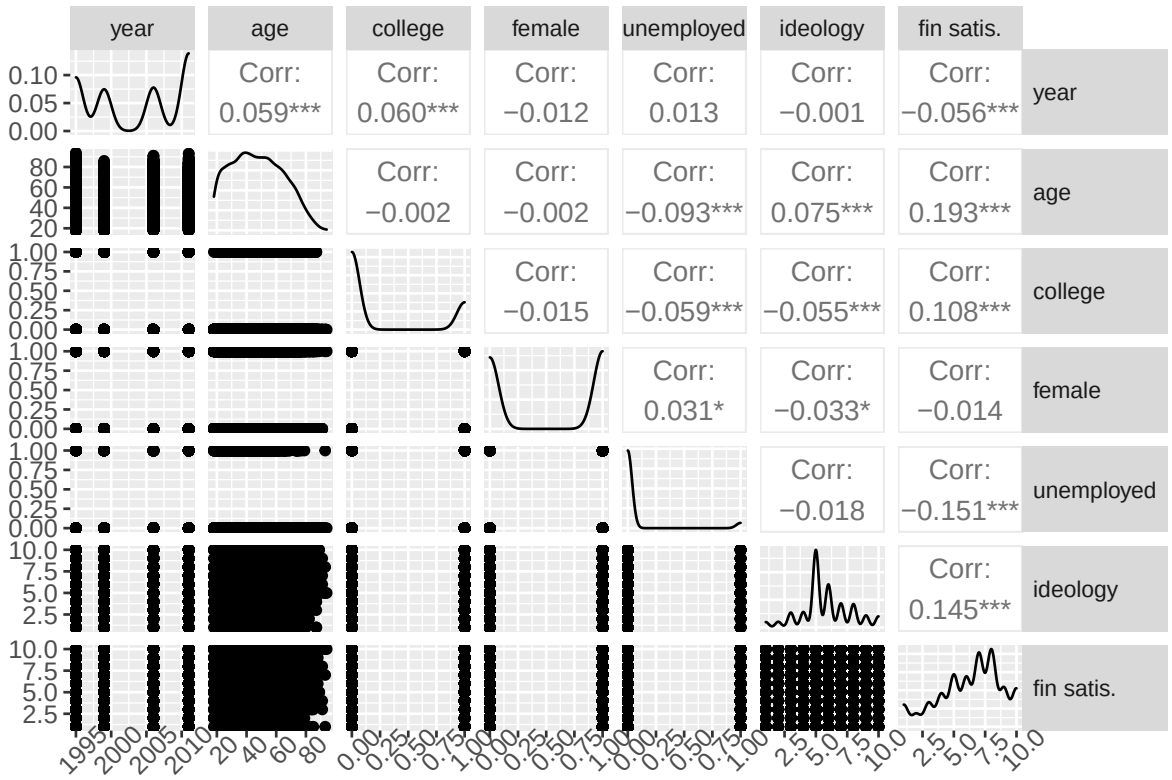



```
ggplot(abortion_data, aes(x = nationalpride, y = aj)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

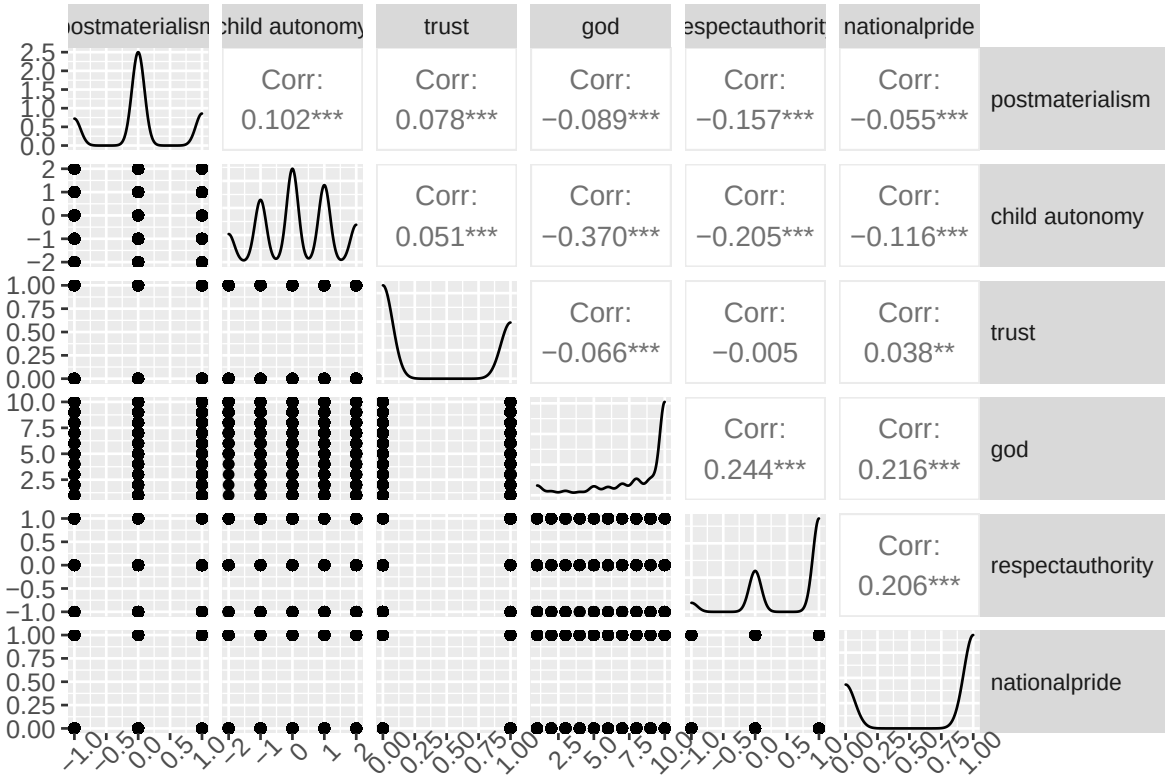


We now test if any of the predictors are strongly correlated with each other.

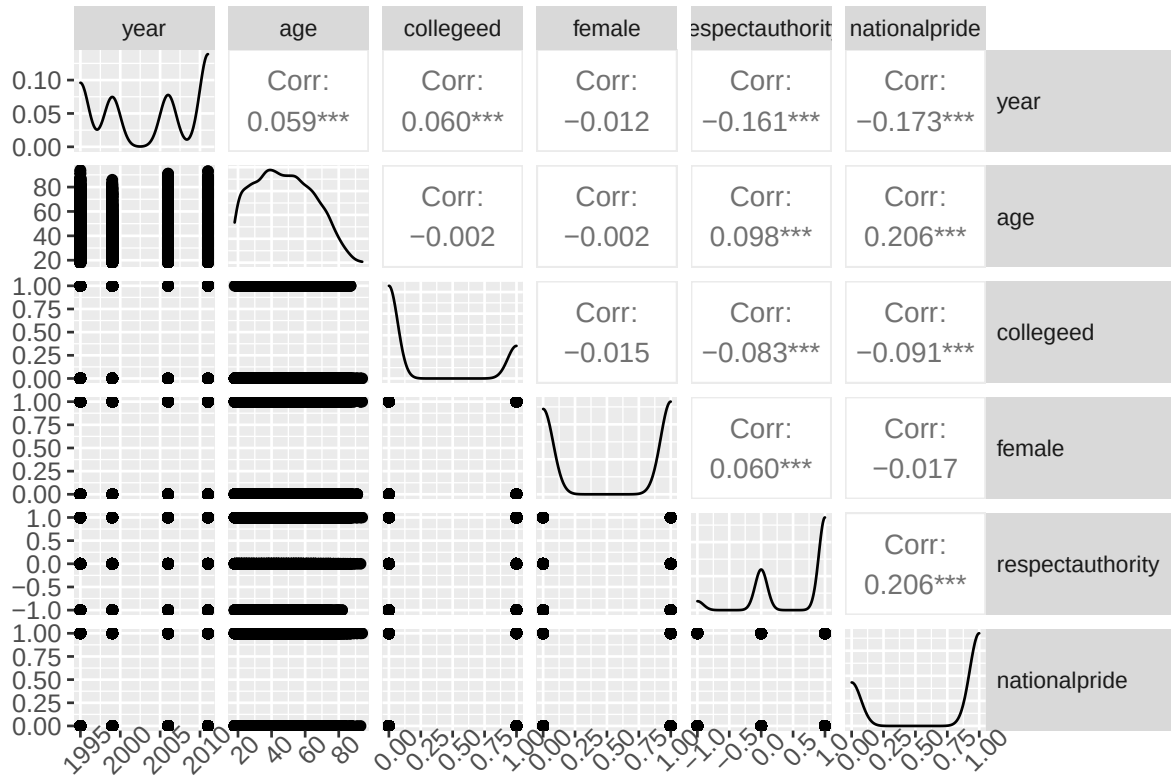
```
#I am doing ggpairs on 13 predictor variables.
#To ensure they fit into the screen, I will make 4 matrices with subsets

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)
    )
```



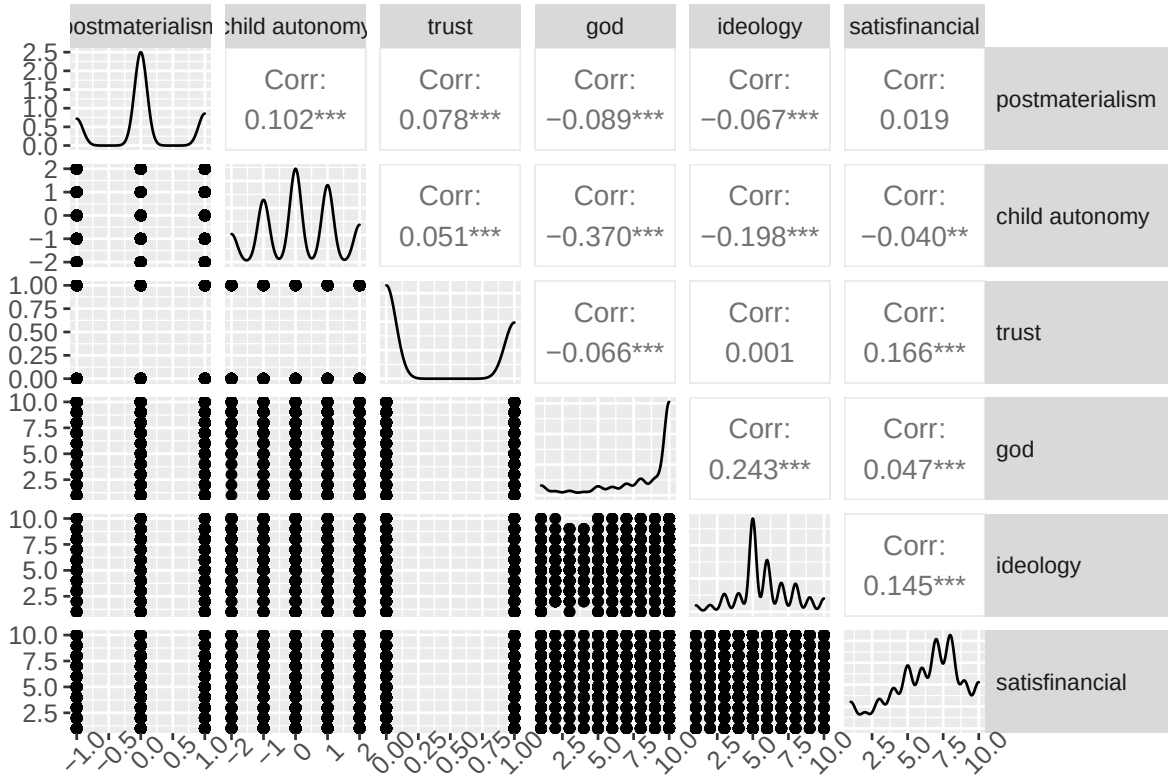

```
ggpairs(abortion_data,
  columns = c("postma4", "cai", "trustmostpeople", "godimportant",
              "respectauthority","nationalpride"),
  columnLabels = c("postmaterialism", "child autonomy", "trust", "god",
              "respectauthority","nationalpride")) +
   theme(
    axis.text.y = element_text(size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