written_assignment

Mohamed Ismael, Mohamed Elgendy

2022-09-20

Background:

The QoG Standard Dataset includes over 2,000 variables, while the QoG Basic Dataset includes approximately 300 variables from the Standard Dataset. The QoG OECD Dataset includes data on OECD member countries and has high data coverage. Each variable entry in this codebook specifies in which dataset the variable can be found. The variables in the Standard, Basic, and OECD datasets are categorized in 19 thematic categories. This categorization should be seen as a guideline rather than a definite classification. Most variables belong only to one category, but some variables belong to more than one category. We have explored some chosen individual variables from this data set to better understand it and then we have compared them against each other to understand if there are correlations/relationships between them. We focused on the effect of liberal democracy on some of the continuous variables, such as political corruption, equal opportunity and freedom of speech. Then, we used linear regression models to analyse the correlations Moreover, we created the multivariate linear regression model, which is a statistical method that is used to model the linear relationship between a dependent variable and multiple independent variables. It is used to make predictions about the dependent variable based on the values of the independent variables.

R Markdown

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Tidy

Import

```
qog_std <- read.csv("dataset/qog_std_cs_jan22.csv")
```

Friendly variable names

These are the variables we would like to focus on.

liberal democracy index (vdem_libdem)

political corruption (vdem_corr)

Freedom of Expression and Belief (fh_feb)

Equal opportunity (bti_eo)

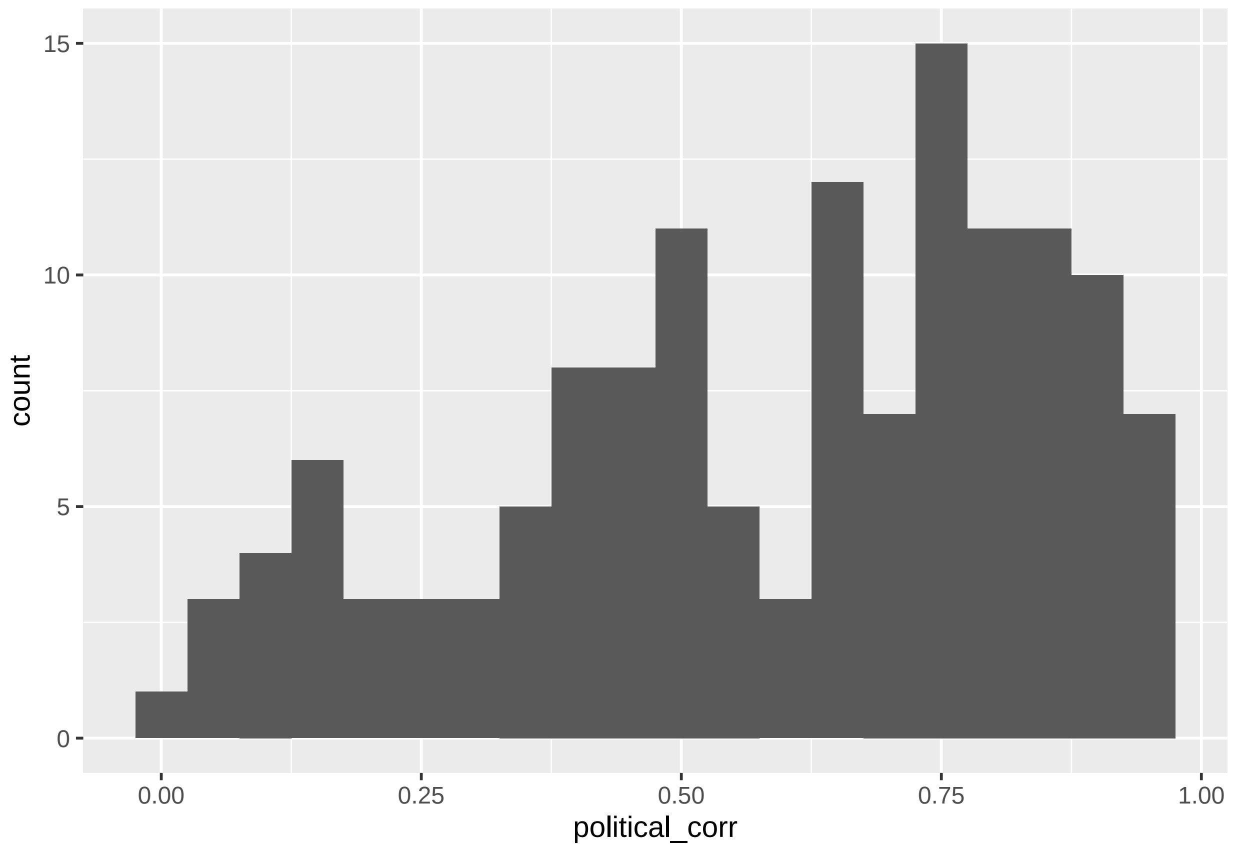
Economic erformance (bti_ep)

```
democracy_effect <- qog_std %>%
  select(
    vdem_libdem,
    vdem_corr,
    bti_eo,
    fh_feb,
    bti_ep
  )%>%
  rename(
    liberal_democracy = vdem_libdem,
    political_corr = vdem_corr,
    equal_oppor = bti_eo,
    freedom_of_speech = fh_feb,
    economic_performance =  bti_ep
  ) %>% drop_na(liberal_democracy,
                political_corr,
                equal_oppor,
                freedom_of_speech
  )
```

We filtered the dataset so we only have the variables we would like to use. We gave them more clear names and dropped all rows with NA values.

Political corruption

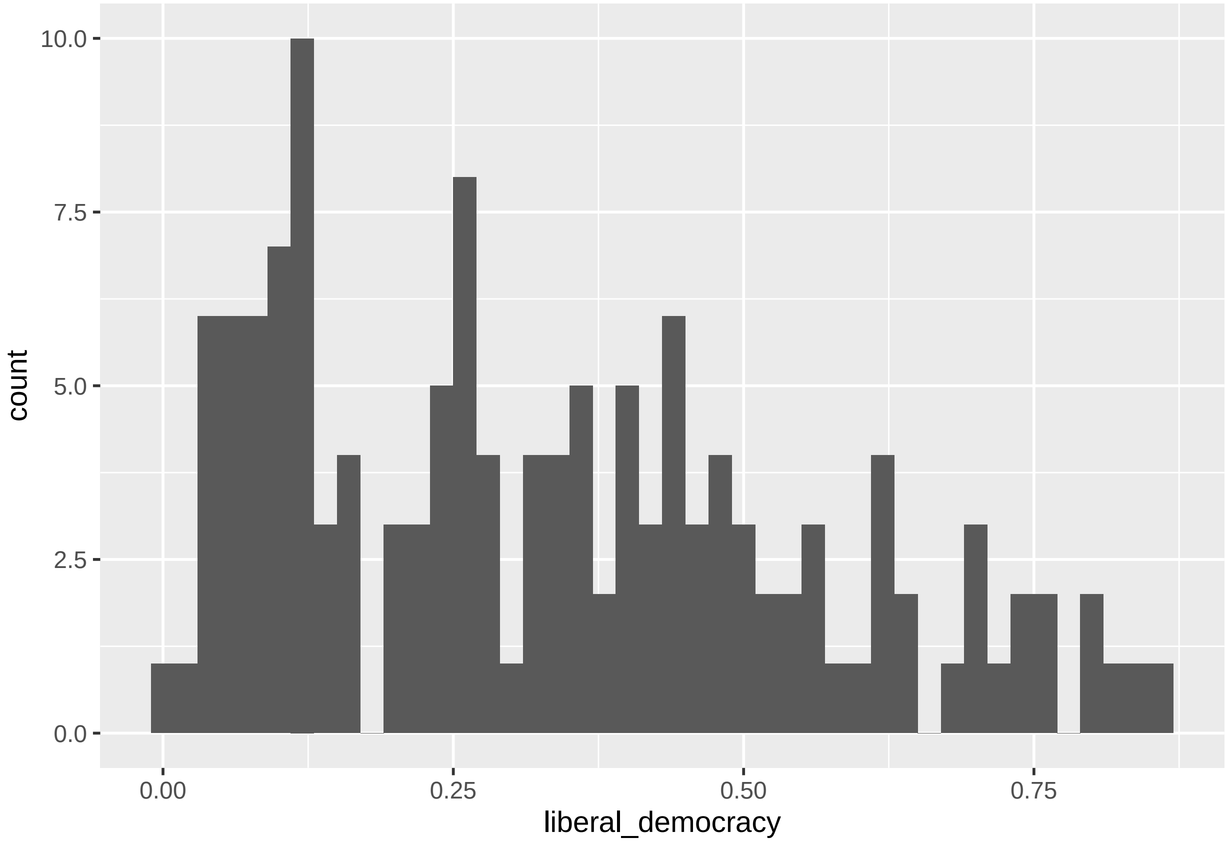
```
  ggplot(data = democracy_effect) +
  geom_histogram(mapping = aes(x = political_corr), binwidth = 0.05)
```



We can see there are more countries in the data set with a higher political corruption index. We can also see that there is a little concentration in the middle with countries that score about average on the political corruption scale.

Liberal democracy

```
ggplot(data = democracy_effect) +
geom_histogram(mapping = aes(x = liberal_democracy), binwidth = 0.02)
```



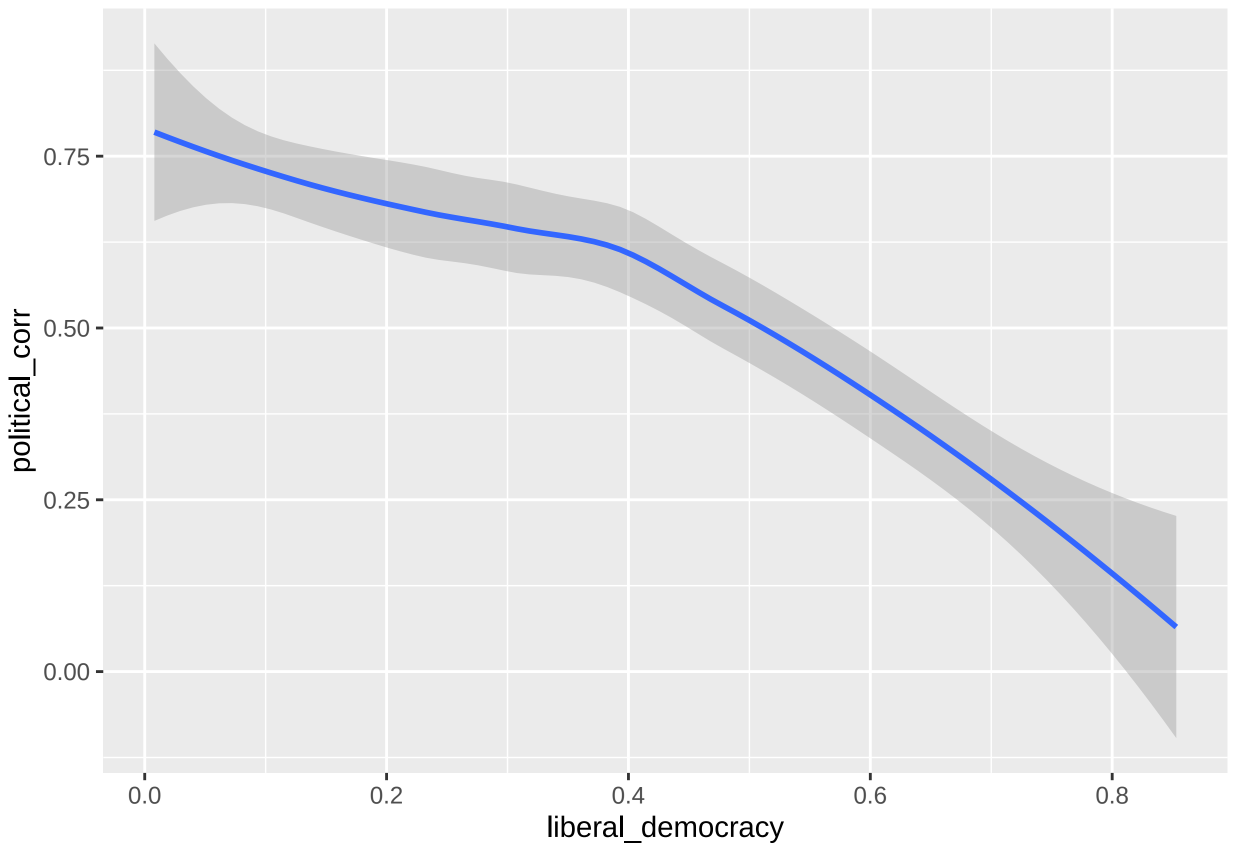
We can see in the histogram for liberal democracy that the higher the value is, the lower the likelihood is for a country to qualify for it.

With a lower binwidth, we are able to more accurately spot a downward thread in the histogram.

Liberal democracy and political corruption

```
ggplot(data = democracy_effect) +
geom_smooth(mapping = aes(x = liberal_democracy, y = political_corr))
```

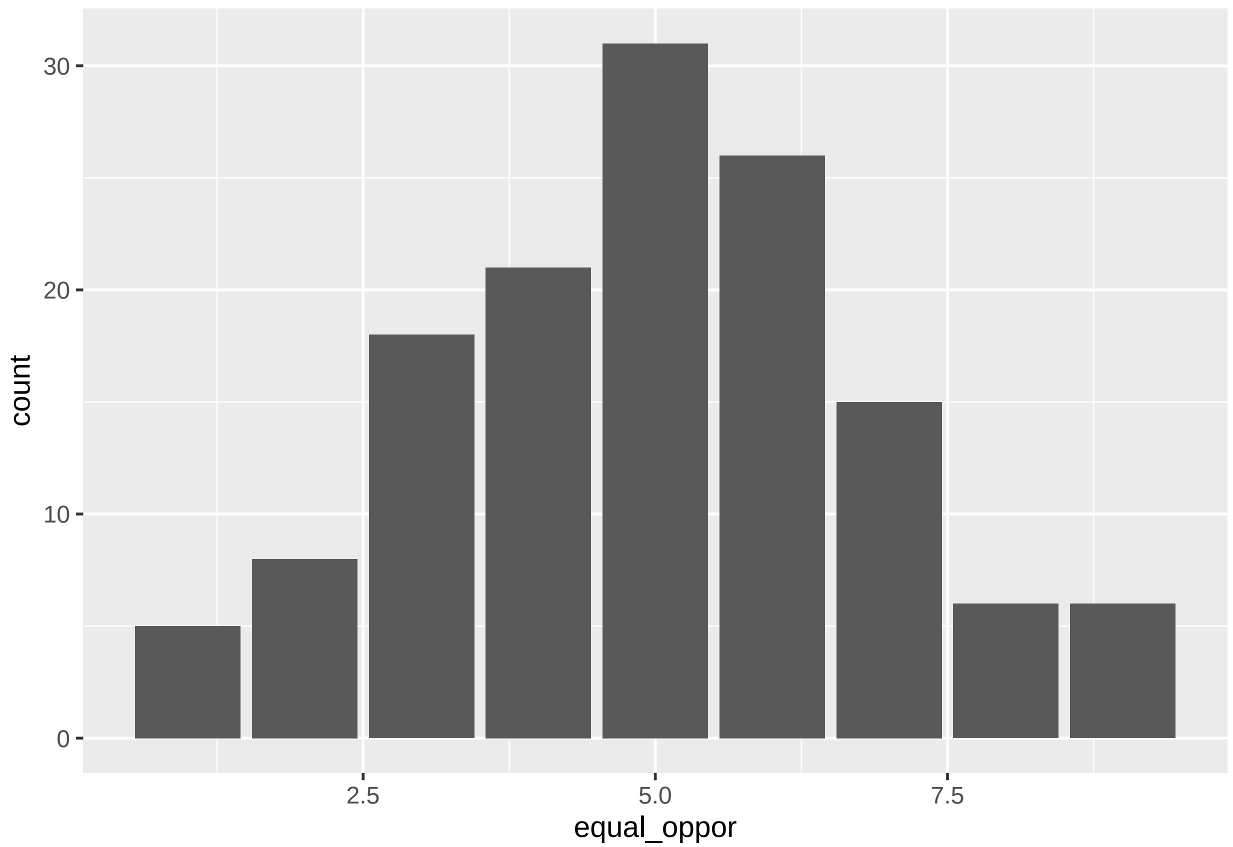
```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



We can see that there is a negative correlation between political corruption and liberal democracy. This is to be expected, because for a country to score high on the liberal democracy index, that country should score low on political corruption. On of the criteria for a liberal democracy is effective checks and balances, and that is not achieved when there is a high amount of political corruption.

Equal opportunity

```
ggplot(data = democracy_effect) +
geom_bar(mapping = aes(x = equal_oppor))
```

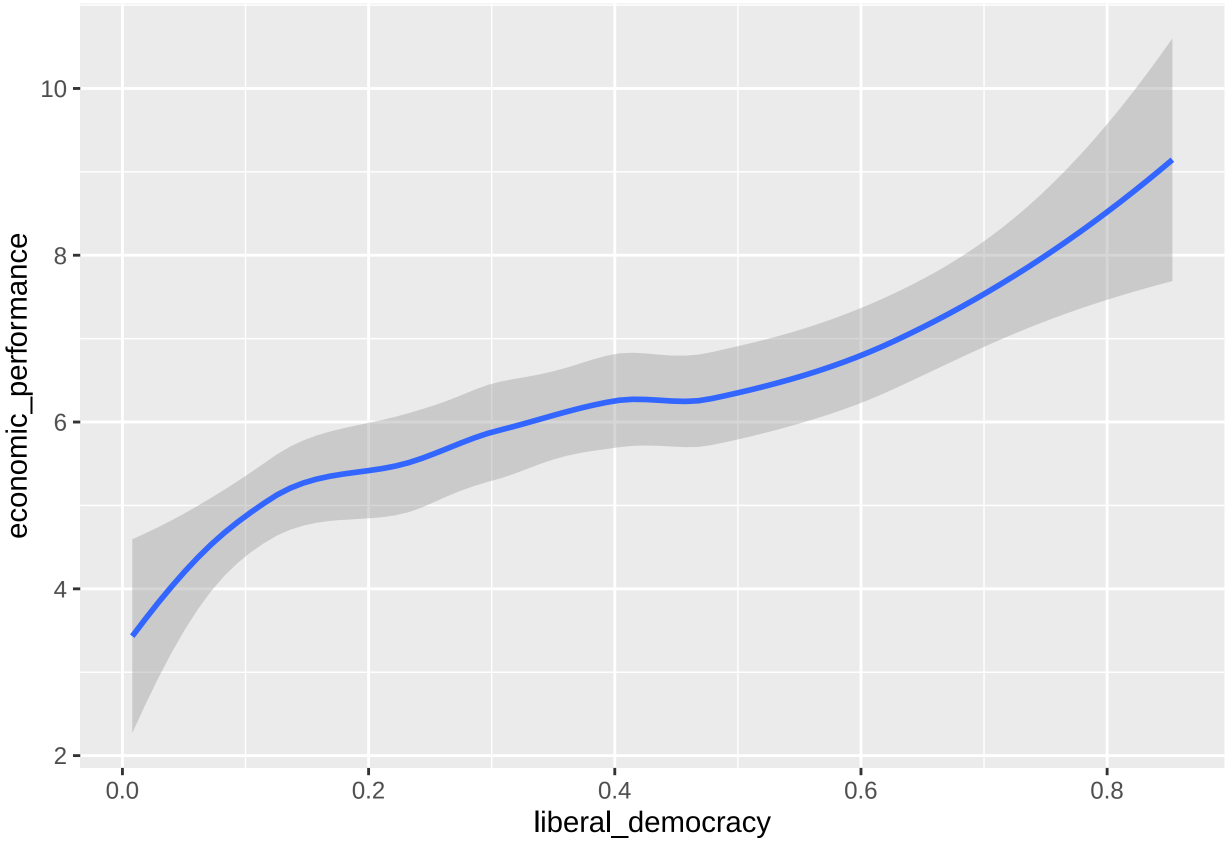


We can see that there is a negative curvilinear relationship between the equal opportunity and the way countries score on it. There are less countries on the low end and high end on the equal opportunity scale as opposed to countries that score between 3 and 7. The highest amount of countries have score of 5.

Liberal democracy and Economic Performance

```
ggplot(data = democracy_effect) +
geom_smooth(mapping = aes(x = liberal_democracy, y = economic_performance))
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```

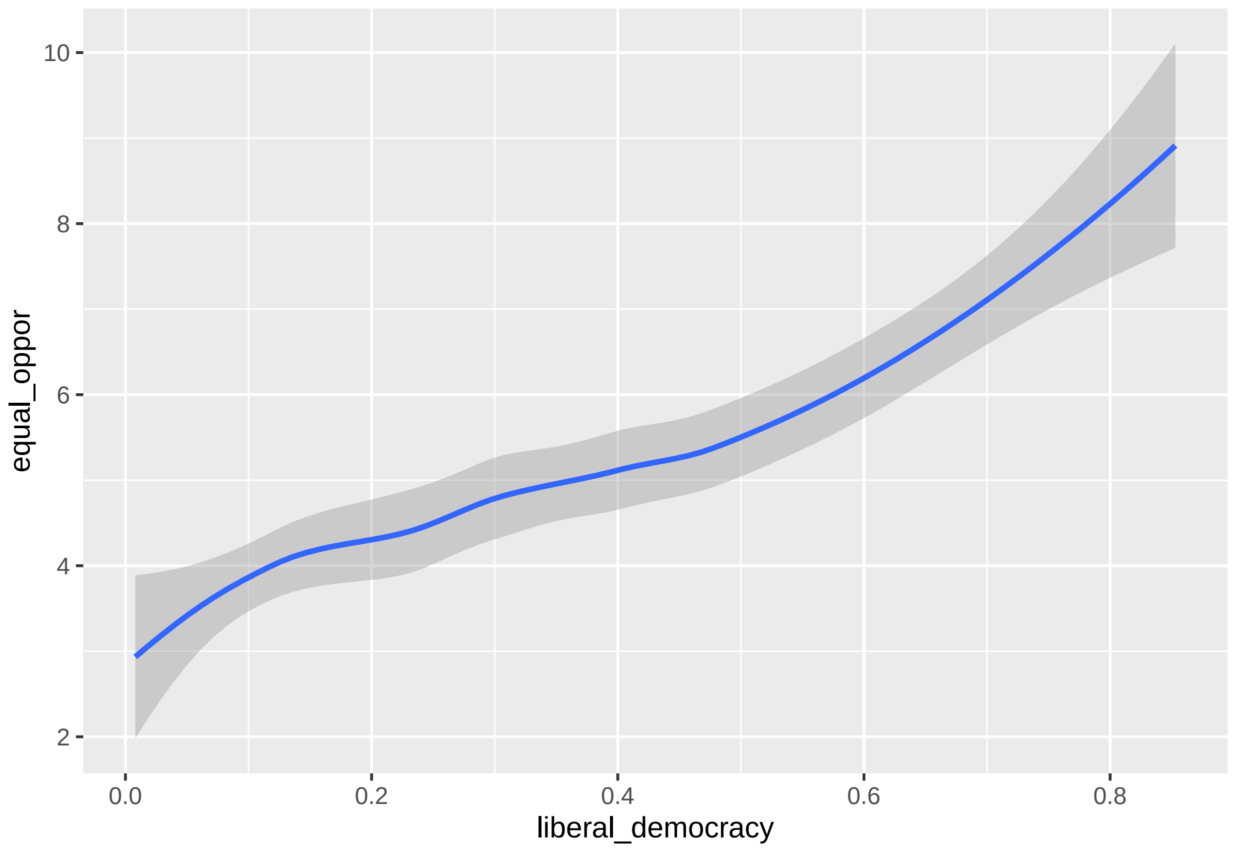


As we can see here that the more liberal democratic the state is, the higher the chance that it will have better economic performance.

Liberal democracy and equal opportunity

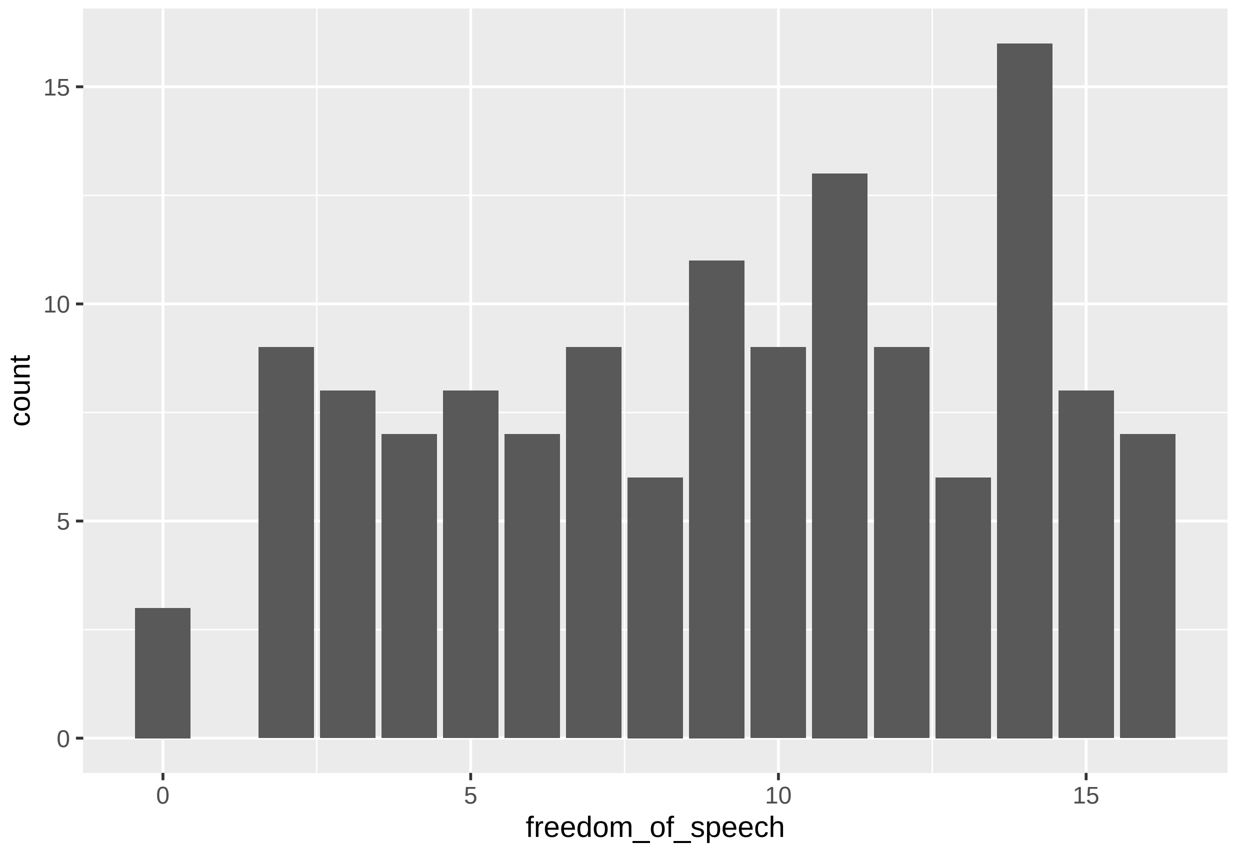
```
ggplot(data = democracy_effect) +
geom_smooth(mapping = aes(x = liberal_democracy, y = equal_oppor))
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



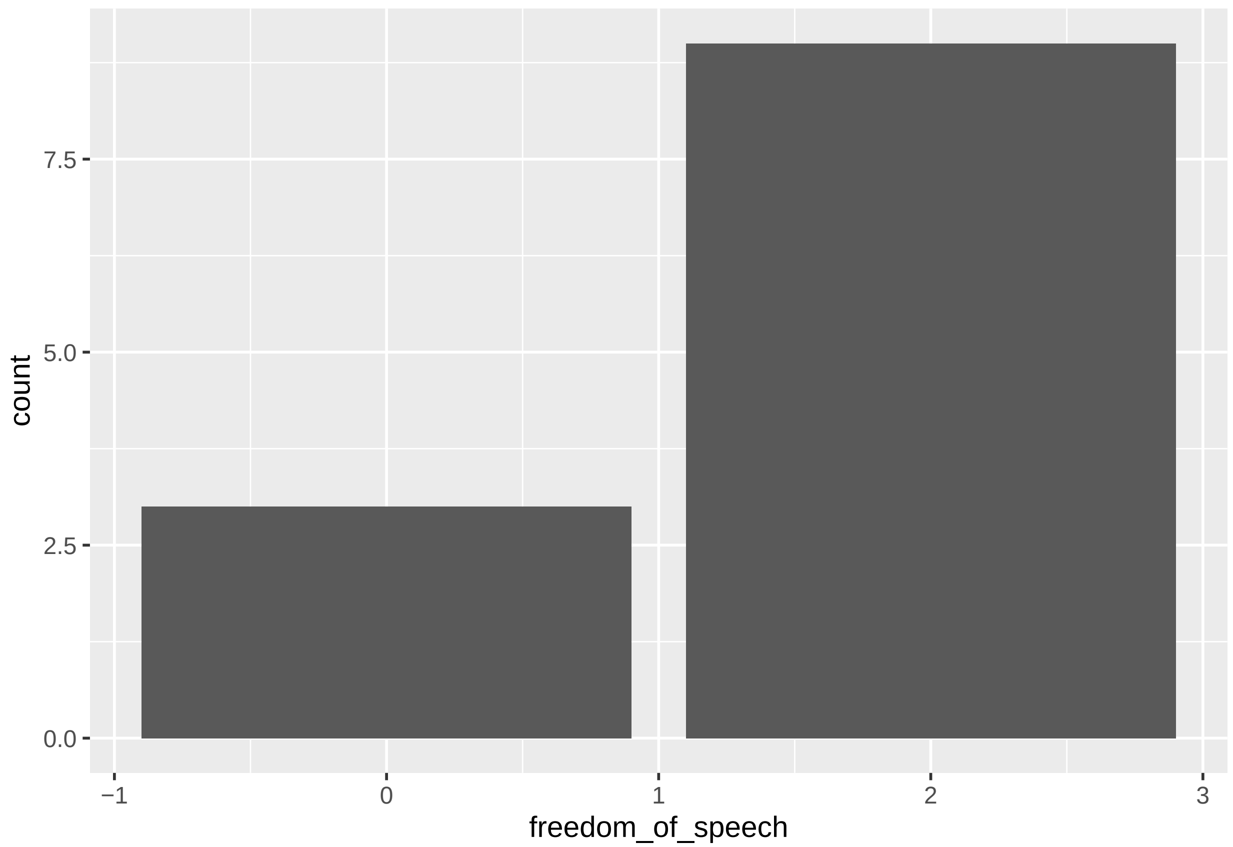
Freedom of speech

```
ggplot(data = democracy_effect) +
geom_bar(mapping = aes(x = freedom_of_speech))
```



Zoom in on low score freedom of speech

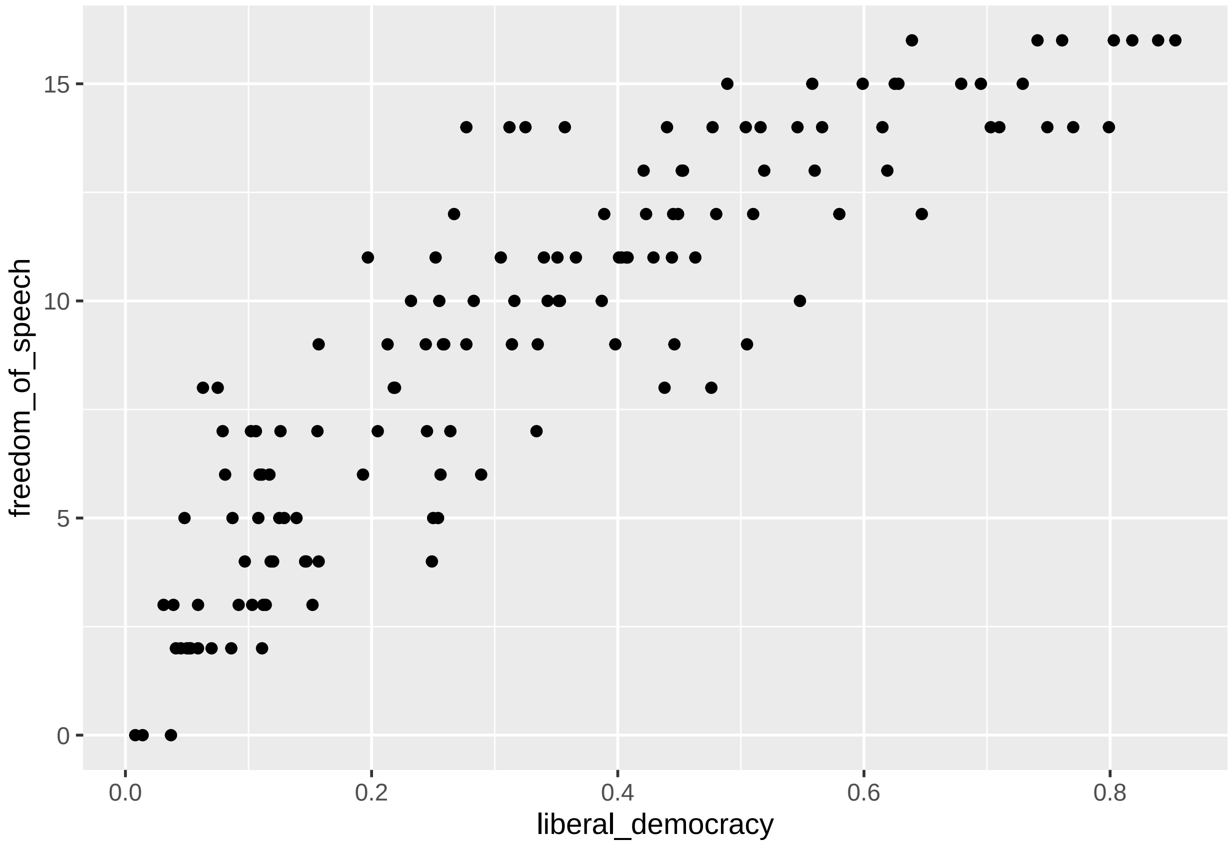
```
freedom_low <- democracy_effect %>%
                filter(freedom_of_speech < 3)

ggplot(data = freedom_low) +
geom_bar(mapping = aes(x = freedom_of_speech))
```



It appears that there's no column in the freedom of speech variable that has the value 1. The bar plot for freedom of speech has a low number of countries with a score lower than 2.

Freedom of speech and liberal democracy

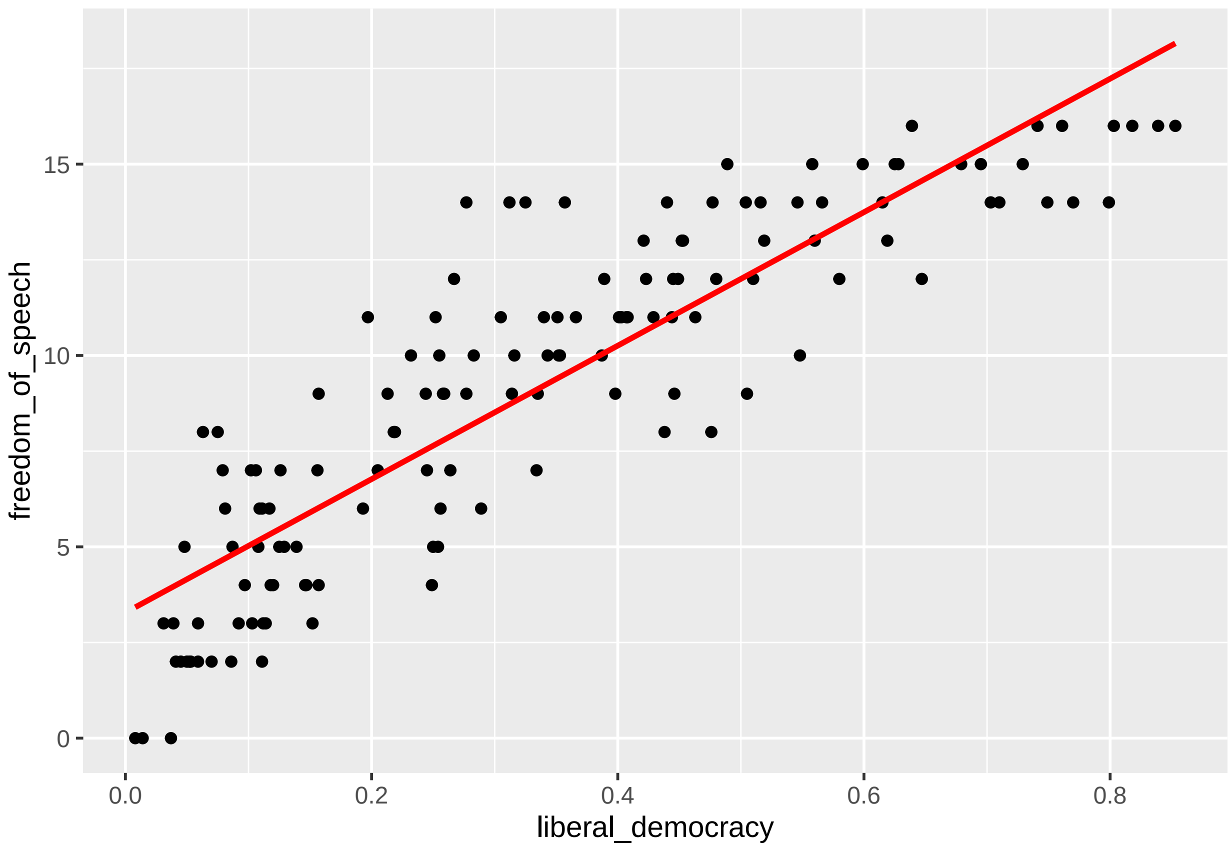
```
ggplot(data = democracy_effect) +
geom_point(mapping = aes(x = liberal_democracy, y = freedom_of_speech))
```



Although the relationship is not too strong, we can see that there is a positive correlation between liberal democracy and freedom of speech. This lines up with our expectations, because for a country to score high on the liberal democracy index, they have to protect civil liberties. A country with a lower score should also score lower on freedom of speech.

```
ggplot(data = democracy_effect, mapping = aes(x = liberal_democracy, y = freedom_of_speech)) +
geom_point() +
geom_smooth(method = lm, color ="red", se=FALSE)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



=

```
  lmspeech <- lm(liberal_democracy ~ freedom_of_speech, data = democracy_effect)

summary(lmspeech)
```

```
##
## Call:
## lm(formula = liberal_democracy ~ freedom_of_speech, data = democracy_effect)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.274626 -0.067883 -0.002132  0.064618  0.247374
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.071346   0.020880  -3.417 0.000839 ***
## freedom_of_speech  0.044498   0.002065  21.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1063 on 134 degrees of freedom
## Multiple R-squared:  0.776,  Adjusted R-squared:  0.7743
## F-statistic: 464.2 on 1 and 134 DF,  p-value: < 2.2e-16
```

Mohamed Ismael liberal democracy & freedom of speech

Mohamed Elgendy liberal democracy & equal opportunity

The following the is the code for creating the regression model:

```
model <- lm(equal_oppor ~ liberal_democracy, data = democracy_effect)
summary(model)
```

```
##
## Call:
## lm(formula = equal_oppor ~ liberal_democracy, data = democracy_effect)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9824 -1.0472 -0.3048  1.2720  3.3284
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.1737     0.2242  14.153   <2e-16 ***
## liberal_democracy   5.4117     0.5592   9.678   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.453 on 134 degrees of freedom
## Multiple R-squared:  0.4114, Adjusted R-squared:  0.407
## F-statistic: 93.67 on 1 and 134 DF,  p-value: < 2.2e-16
```

Based on that outcome, the model is trying to predict the dependent variable equal_oppor based on the independent variable liberal_democracy.

The "Residuals" section shows the difference between the predicted values and the actual values for the dependent variable. The "Coefficients" section shows the regression coefficients for each independent variable in the model. In this case, there is only one independent variable, liberal_democracy, and the coefficient is 5.4117. This means that for every one unit increase in liberal_democracy, we would expect to see a 5.4117 unit increase in equal_oppor. The intercept is 3.1737, which is the predicted value of equal_oppor when liberal_democracy is zero.

In conclusion:

1. Residual standard error: The residual standard error is a measure of the amount of error in the model. A lower residual standard error indicates a better fit. In this case, the residual standard error is 1.453, which is relatively low. This suggests that the model is a good fit for the data.

2. R-squared values: The R-squared values are measures of how well the model fits the data. The R-squared value is the proportion of the variance in the dependent variable that is explained by the independent variable. A value closer to 1 indicates a better fit. In this case, the R-squared values are 0.4114 and 0.407, which are relatively high. This suggests that the model explains a significant amount of the variance in the data.

3. p-value for the F-statistic: The p-value for the F-statistic is used to test the overall significance of the model. A low p-value (typically less than 0.05) indicates that the model is a good fit for the data. In this case, the p-value is extremely low (less than 2.2e-16), indicating that the model is a good fit for the data.

```
model <- lm(equal_oppor ~ liberal_democracy + freedom_of_speech + political_corr + economic_performance

summary(model)
```

```
##
## Call:
## lm(formula = equal_oppor ~ liberal_democracy + freedom_of_speech +
##     political_corr + economic_performance, data = democracy_effect)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.06668 -0.85063  0.02121  0.93788  2.59134
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.20498    0.63722   5.030 1.58e-06 ***
## liberal_democracy     1.00546    1.24626   0.807 0.421258
## freedom_of_speech     0.08116    0.05478   1.481 0.140891
## political_corr       -2.05663    0.60709  -3.388 0.000931 ***
## economic_performance  0.32486    0.06374   5.097 1.18e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.235 on 131 degrees of freedom
## Multiple R-squared:  0.5848, Adjusted R-squared:  0.5721
## F-statistic: 46.13 on 4 and 131 DF,  p-value: < 2.2e-16
```

This multivariate linear regression model is analyzing the relationship between the outcome variable "equal_oppor" and four predictor variables: "liberal_democracy", "freedom_of_speech", "political_corr", and "economic_performance". The model estimates the following coefficients for each predictor variable:

1. A one unit increase in "liberal_democracy" is associated with a 1.00546 unit increase in "equal_oppor", holding all other variables constant.
2. A one unit increase in "freedom_of_speech" is associated with a 0.08116 unit increase in "equal_oppor", holding all other variables constant.
3. A one unit increase in "political_corr" is associated with a 2.05663 unit decrease in "equal_oppor", holding all other variables constant.
4. A one unit increase in "economic_performance" is associated with a 0.32486 unit increase in "equal_oppor", holding all other variables constant.

The model's residual standard error is 1.235 and the multiple R-squared value is 0.5848, indicating that the model explains approximately 58.5% of the variance in the outcome variable. The F-statistic of 46.13 and a p-value of < 2.2e-16 indicate that the model is a good fit for the data. The p-values for the coefficients of "liberal_democracy" and "freedom_of_speech" are relatively high, indicating that the relationship between these variables and the outcome variable is not statistically significant at a significance level of 0.05. However, the p-values for the coefficients of "political_corr" and "economic_performance" are lower, indicating that the relationship between these variables and the outcome variable is statistically significant at a significance level of 0.05.