Data Exploration

Group 4

2023-07-06

```
# loads all needed libraries
library(dplyr)
library(tidyr)
library(readxl)
library(tidyverse)
library(reshape2)
```

Exploring the Data

What is your outcome variable(s)? How well does it measure the outcome you are interested in? How does it relate to your hypotheses? What are your key explanatory variables? In addition, create a table of summary statistics for the variables you are planning to use.

General Answer: Our outcome variable is civil society participation, which is v2x_cspart in the code.

We will use the predictor variables listed above, which include: civil society repression effort scores, v2csreprss social support scores, happiness freedom to make life choices, happiness presence of war, e_civil_war presence of coups, e_pt_coup participation rate in civil society, v2x_partipdem generosity, happiness education, e_peaveduc and government corruption. v2x_corr

Merging data - James

```
wh_2023 <- read_excel("/cloud/project/data/wh_2023.xls")
# View(wh_2023)

load("/cloud/project/data/vdemdata-master/data/vdem.RData")
```

```r
# colnames(wh_2023)
# colnames(vdem)
```

```r
need_vdem=c("country_name","year","v2csreprss",'e_civil_war',"e_pt_coup",
"v2x_partipdem","e_peaveduc","v2x_corr","v2x_cspart","v2xcs_ccsi")

vdem_use=vdem[,need_vdem]

need_wh=c("Country name","year","Social support","Freedom to make life choices","Generosit
wh_use=wh_2023[,need_wh]

colnames(wh_use)[1] <- "country_name"

total=merge(vdem_use, wh_use, by = c("country_name", "year"),
            all.x=TRUE,
            all.y=TRUE)
```

Cleaning data - Alicia

Background on the new data set

Because there were too many NA values in each of the column of the data frame, in order to ensure the machine learning approaches run properly, we need to either omit these values or assign a certain mean/median or mode value to replace the missing values here. We tried to omit the NA values first but that would leave us with only 73 out of over 27000 obs. So I used the median as the replacement value for the NAs in each of the columns in the data set.

The exception variable is the e_civil_war, which is binary data. The data is also skewed with the majority of obs 0 or NA. It is thus better to use the median or mode instead of mean. However, this might still not be the best way to deal with missing value. Our other options are to not use this variable in the model or use mice() package to deal with missing values.

We choose only data from 2012 onward for our analyses. Also, when we do data visualizations, we see that there are a lot of outliers in certain columns. We will therefore use bestNormalize() to clean the data in these columns.

We did not have to make any new variables. We only have to merge the two data sets as executed above.

```r
# Select the time range from 2012 (inclusive) and later only
total_2012 <- total[total$year > 2011,]
head(total_2012)
```

```
    country_name year v2csreprss e_civil_war e_pt_coup v2x_partipdem e_peaveduc
224  Afghanistan 2012      1.929          NA         0         0.151      3.813
225  Afghanistan 2013      1.269          NA         0         0.158      3.813
226  Afghanistan 2014      1.269          NA         0         0.157      3.813
227  Afghanistan 2015      1.152          NA         0         0.165      3.813
228  Afghanistan 2016      1.425          NA         0         0.151      3.813
229  Afghanistan 2017      1.205          NA         0         0.145      3.813
    v2x_corr v2x_cspart v2xcs_ccsi Social support Freedom to make life choices
224    0.945      0.664      0.745      0.5206367                     0.5309350
225    0.922      0.692      0.725      0.4835519                     0.5779554
226    0.918      0.692      0.725      0.5255684                     0.5085140
227    0.891      0.758      0.739      0.5285972                     0.3889276
228    0.893      0.708      0.736      0.5590718                     0.5225662
229    0.886      0.752      0.730      0.4908801                     0.4270109
     Generosity
224  0.23758759
225  0.06266622
226  0.10575488
227  0.08165228
228  0.04391602
229 -0.11941047
```

```r
# changing the NA entries with a median of that column
summary(total_2012)
```

```
 country_name            year        v2csreprss       e_civil_war
 Length:2051        Min.   :2012   Min.   :-3.7590   Min.   : NA
 Class :character   1st Qu.:2014   1st Qu.:-0.2740   1st Qu.: NA
 Mode  :character   Median :2017   Median : 1.1320   Median : NA
                    Mean   :2017   Mean   : 0.8821   Mean   :NaN
                    3rd Qu.:2020   3rd Qu.: 2.0770   3rd Qu.: NA
                    Max.   :2022   Max.   : 3.3240   Max.   : NA
                                   NA's   :82        NA's   :2051
   e_pt_coup       v2x_partipdem      e_peaveduc        v2x_corr
 Min.   :0.0000   Min.   :0.0080   Min.   : 1.310   Min.   :0.002
 1st Qu.:0.0000   1st Qu.:0.1670   1st Qu.: 5.704   1st Qu.:0.219
```

```
 Median :0.0000   Median :0.3240   Median : 8.140   Median :0.543
 Mean   :0.0108   Mean   :0.3353   Mean   : 8.012   Mean   :0.498
 3rd Qu.:0.0000   3rd Qu.:0.4860   3rd Qu.:10.690   3rd Qu.:0.766
 Max.   :2.0000   Max.   :0.8140   Max.   :13.300   Max.   :0.971
 NA's   :476      NA's   :82       NA's   :610      NA's   :82
   v2x_cspart       v2xcs_ccsi     Social support
 Min.   :0.0250   Min.   :0.0080   Min.   :0.2282
 1st Qu.:0.5580   1st Qu.:0.4720   1st Qu.:0.7406
 Median :0.7290   Median :0.7790   Median :0.8321
 Mean   :0.6697   Mean   :0.6687   Mean   :0.8082
 3rd Qu.:0.8450   3rd Qu.:0.9040   3rd Qu.:0.9048
 Max.   :0.9890   Max.   :0.9790   Max.   :0.9873
 NA's   :82       NA's   :82       NA's   :569
 Freedom to make life choices   Generosity
 Min.   :0.3035               Min.   :-0.3375
 1st Qu.:0.6862               1st Qu.:-0.1157
 Median :0.7832               Median :-0.0228
 Mean   :0.7654               Mean   : 0.0005
 3rd Qu.:0.8706               3rd Qu.: 0.0947
 Max.   :0.9852               Max.   : 0.7027
 NA's   :583                  NA's   :597
```

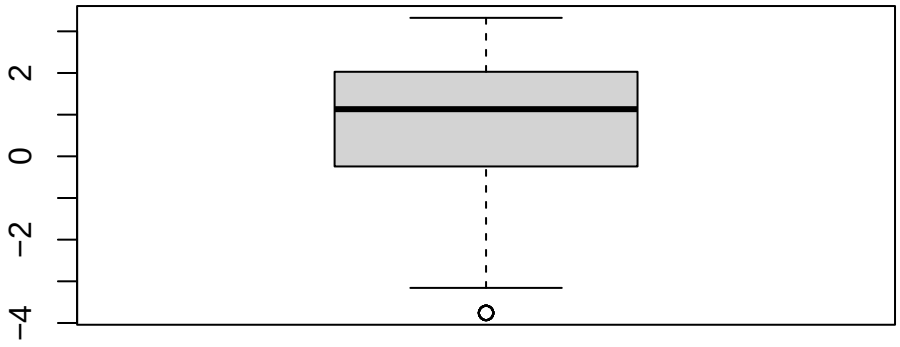
```
total_not_NA <- total_2012 %>% na.omit()
# Because we are missing too many observations when we drop NAs, we will assign
# median of each column to the NA values instead

# We might need a different method for e_civil_war
# because it is all binary 0 and 1
# and it is skewed data

# Get the summary of the e_civil_war variable
total_2012 %>% group_by(e_civil_war) %>% count()
```

```
# A tibble: 1 x 2
# Groups:   e_civil_war [1]
  e_civil_war     n
        <dbl> <int>
1          NA  2051
```

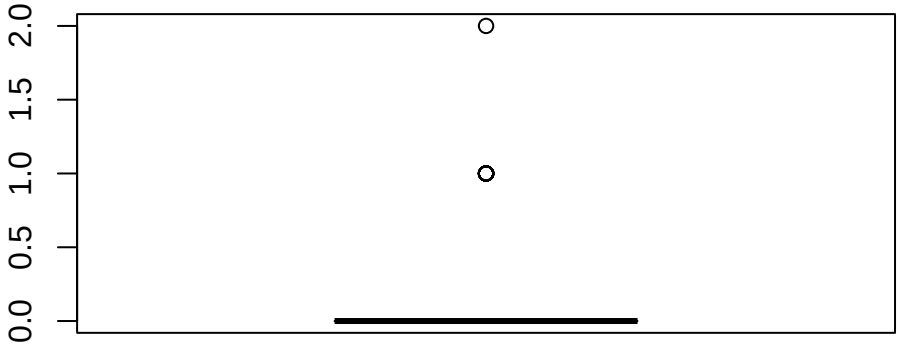
```r
#replace missing values in each numeric column with median value of column
new_total <- total_2012 %>% mutate(across(where(is.numeric),~replace_na(.,median(.,na.rm=T

# Get the summary of our combined data set
summary_table <- summary(new_total)
summary_table
```

```
country_name            year        v2csreprss       e_civil_war
Length:2051        Min.   :2012   Min.   :-3.7590   Min.   : NA
Class :character   1st Qu.:2014   1st Qu.:-0.2430   1st Qu.: NA
Mode  :character   Median :2017   Median : 1.1320   Median : NA
                   Mean   :2017   Mean   : 0.8921   Mean   :NaN
                   3rd Qu.:2020   3rd Qu.: 2.0290   3rd Qu.: NA
                   Max.   :2022   Max.   : 3.3240   Max.   : NA
                                                    NA's   :2051
  e_pt_coup        v2x_partipdem      e_peaveduc       v2x_corr
Min.   :0.000000   Min.   :0.0080   Min.   : 1.31   Min.   :0.0020
1st Qu.:0.000000   1st Qu.:0.1690   1st Qu.: 6.75   1st Qu.:0.2295
Median :0.000000   Median :0.3240   Median : 8.14   Median :0.5430
Mean   :0.008289   Mean   :0.3349   Mean   : 8.05   Mean   :0.4998
3rd Qu.:0.000000   3rd Qu.:0.4820   3rd Qu.: 9.63   3rd Qu.:0.7560
Max.   :2.000000   Max.   :0.8140   Max.   :13.30   Max.   :0.9710

  v2x_cspart       v2xcs_ccsi     Social support
Min.   :0.0250   Min.   :0.0080   Min.   :0.2282
1st Qu.:0.5715   1st Qu.:0.4905   1st Qu.:0.7829
Median :0.7290   Median :0.7790   Median :0.8321
Mean   :0.6720   Mean   :0.6731   Mean   :0.8148
3rd Qu.:0.8375   3rd Qu.:0.9000   3rd Qu.:0.8828
Max.   :0.9890   Max.   :0.9790   Max.   :0.9873

Freedom to make life choices   Generosity
Min.   :0.3035                Min.   :-0.337527
1st Qu.:0.7286                1st Qu.:-0.076792
Median :0.7832                Median :-0.022774
Mean   :0.7705                Mean   :-0.006283
3rd Qu.:0.8349                3rd Qu.: 0.043048
Max.   :0.9852                Max.   : 0.702708
```

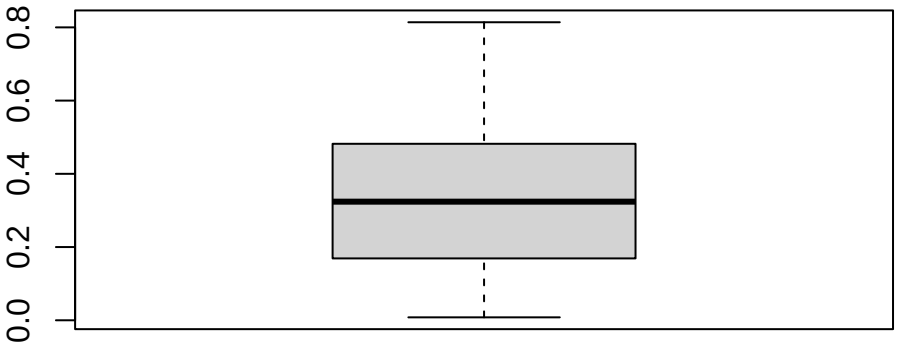
```r
# Boxplots and Histograms
# Draw boxplots for each column

# Choose only appropriate columns to draw boxplots and histograms
# Certain columns will be normalized to be used in regression models later
selected <- new_total %>% select(-c(year, country_name, e_civil_war))

for (column in colnames(selected)) {
  boxplot(selected[,column], xlab = column)
}
```



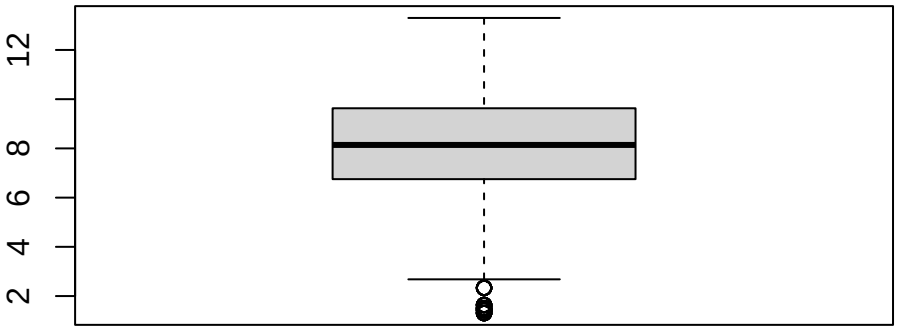
v2csreprss



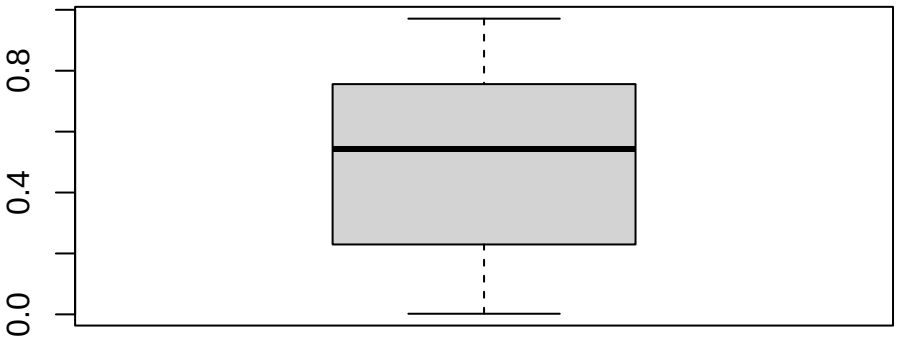
e_pt_coup



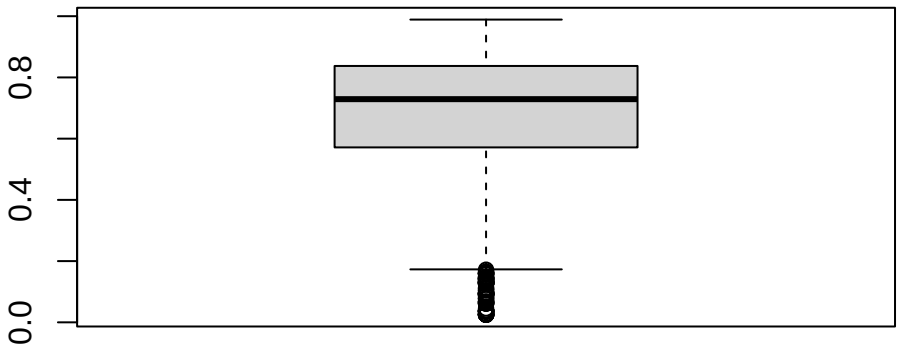
v2x_partipdem



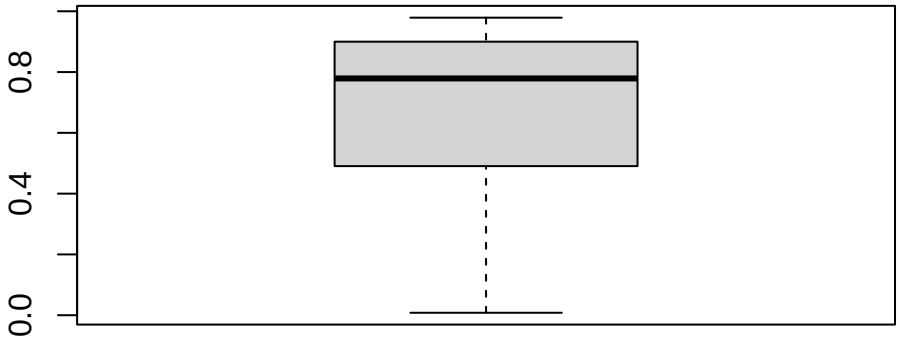
e_peaveduc



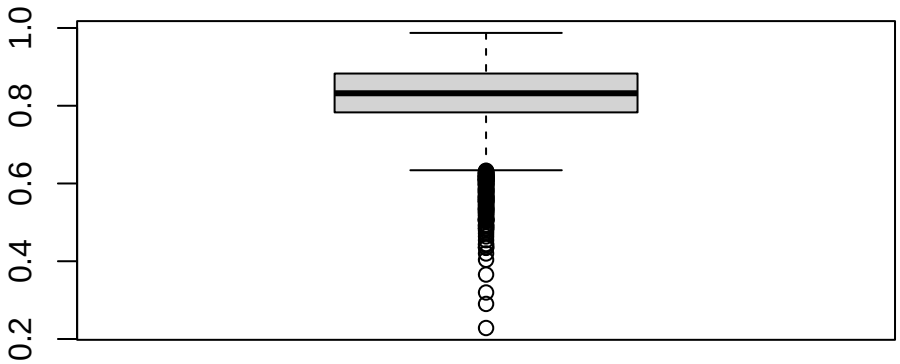
v2x_corr



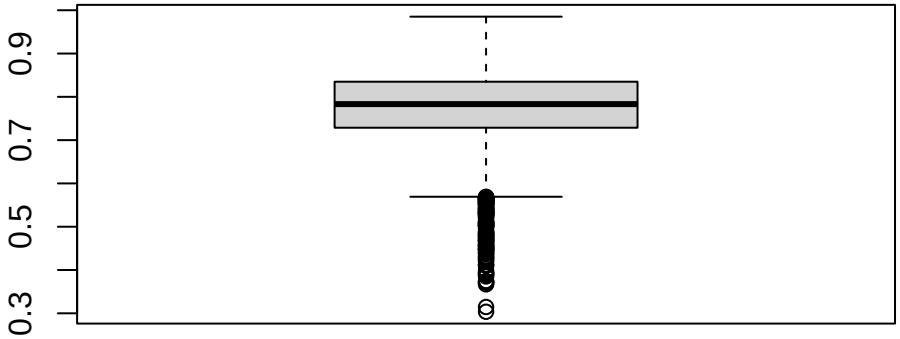
v2x_cspart



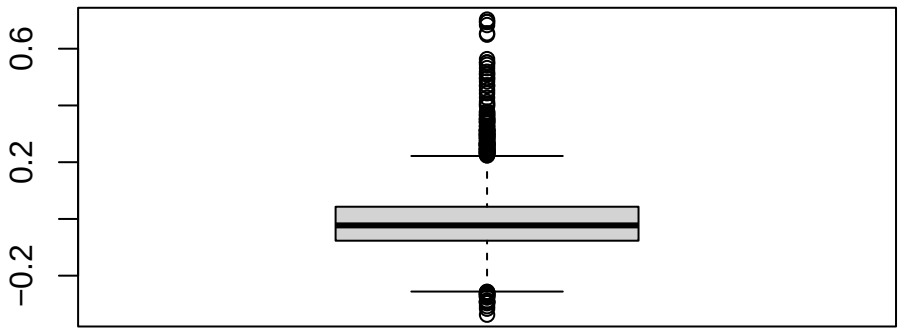
v2xcs_ccsi



Social support

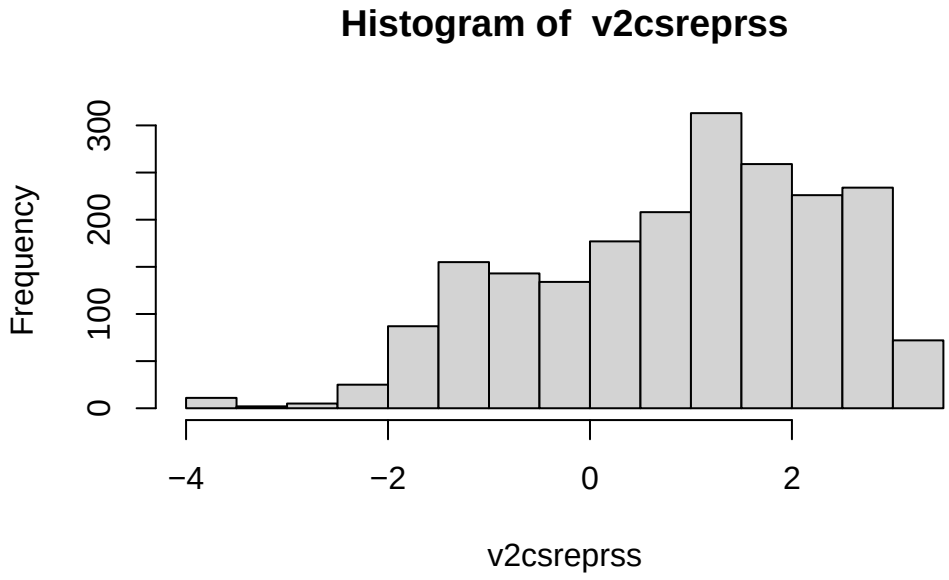


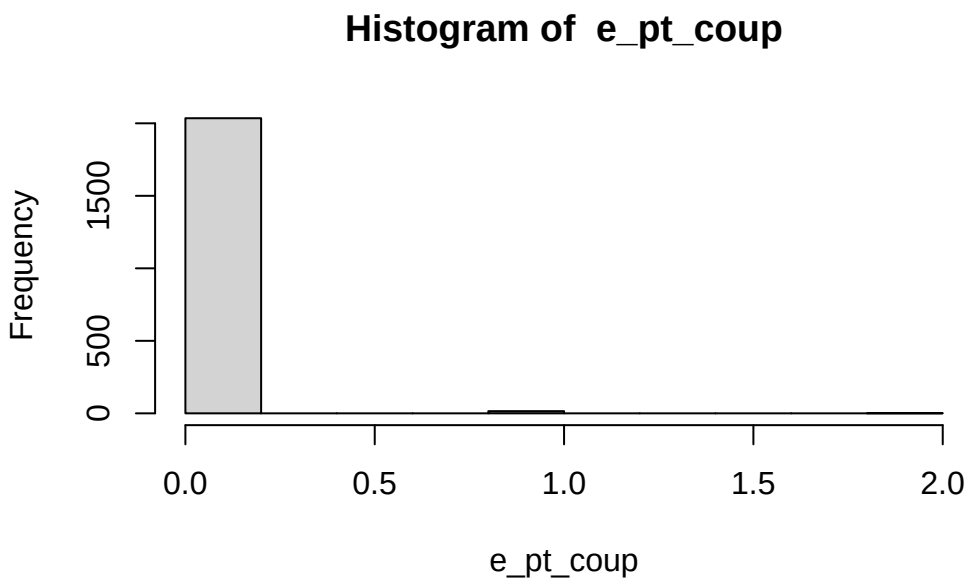
Freedom to make life choices

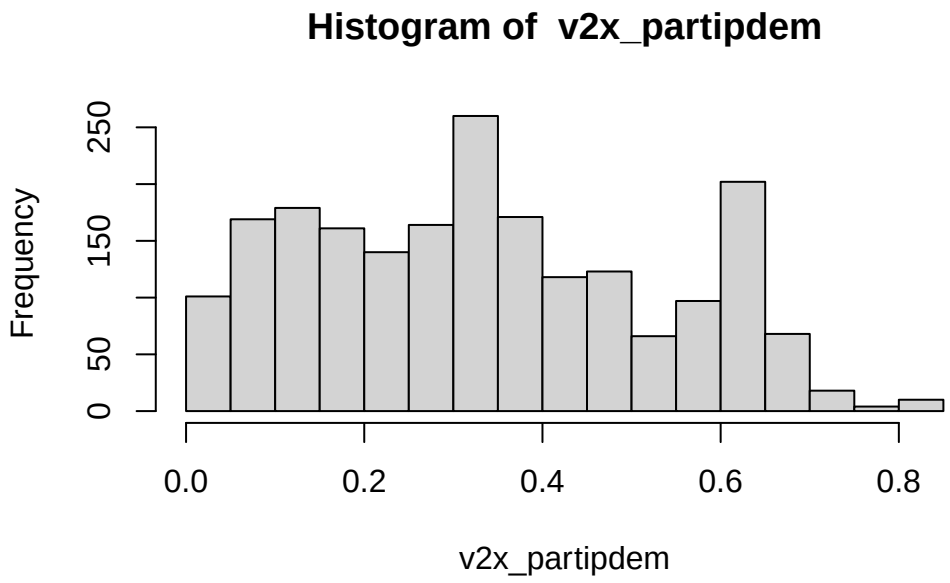


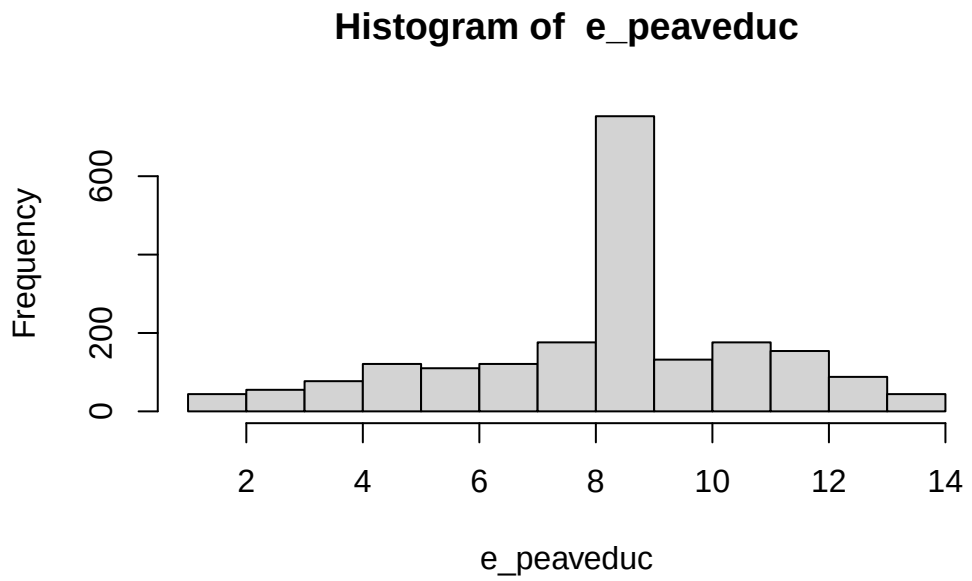
Generosity

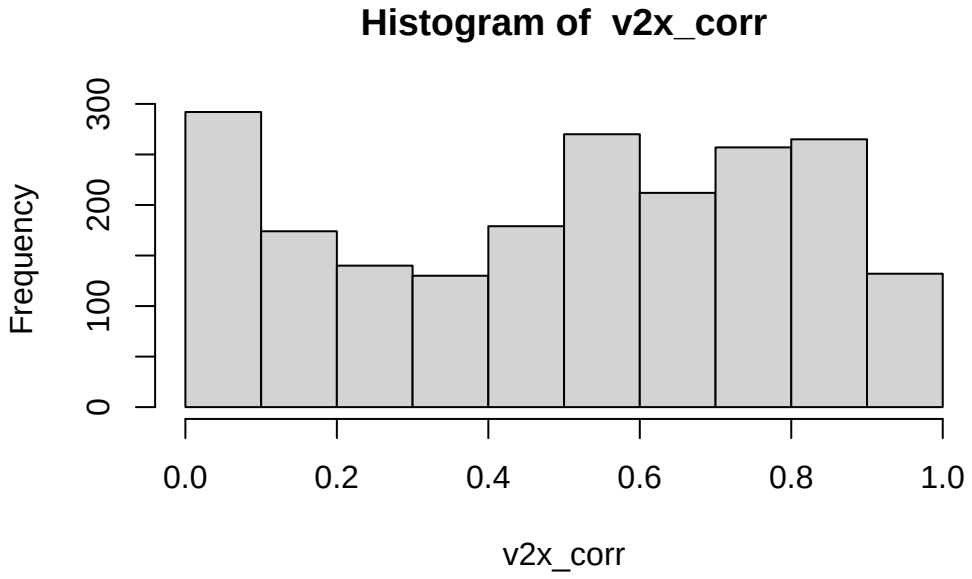
```
# Draw histograms for each column
for (column in colnames(selected)) {
  hist(selected[,column], xlab = column, main = paste("Histogram of ", column))
}
```

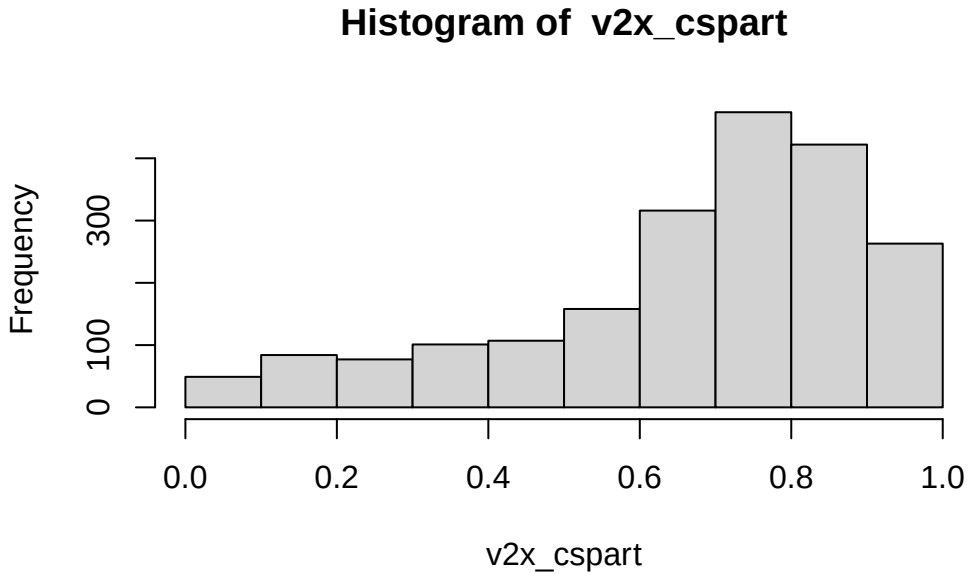


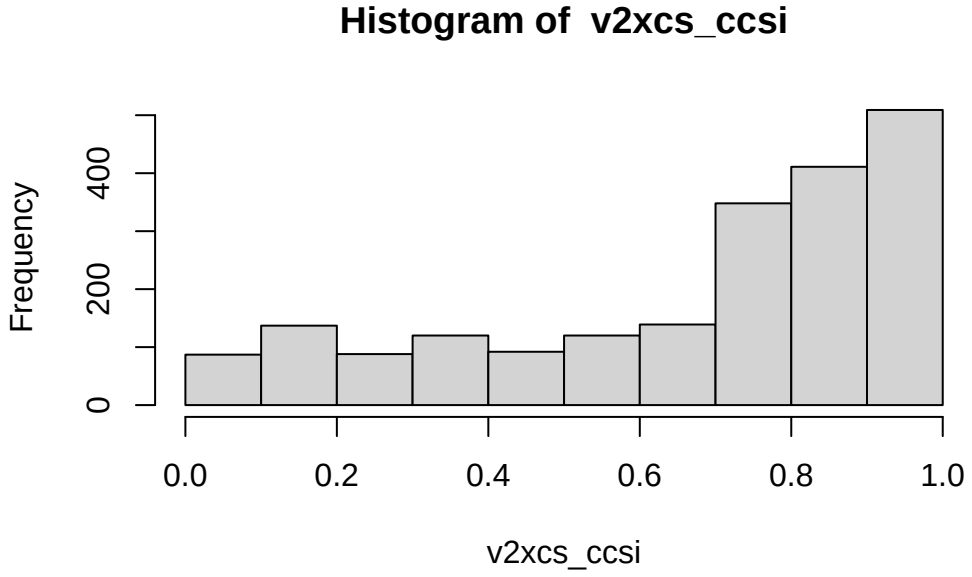

Histogram of e_pt_coup
Frequency
0
500
1500
0.0
0.5
1.0
1.5
2.0
e_pt_coup

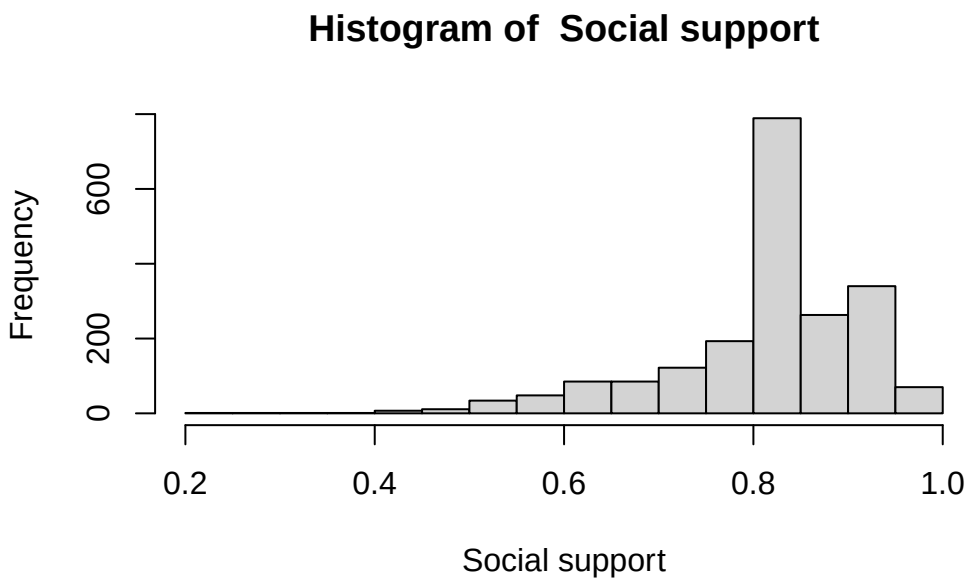

Histogram of v2x_partipdem
Frequency
0
50
150
250
0.0
0.2
0.4
0.6
0.8
v2x_partipdem

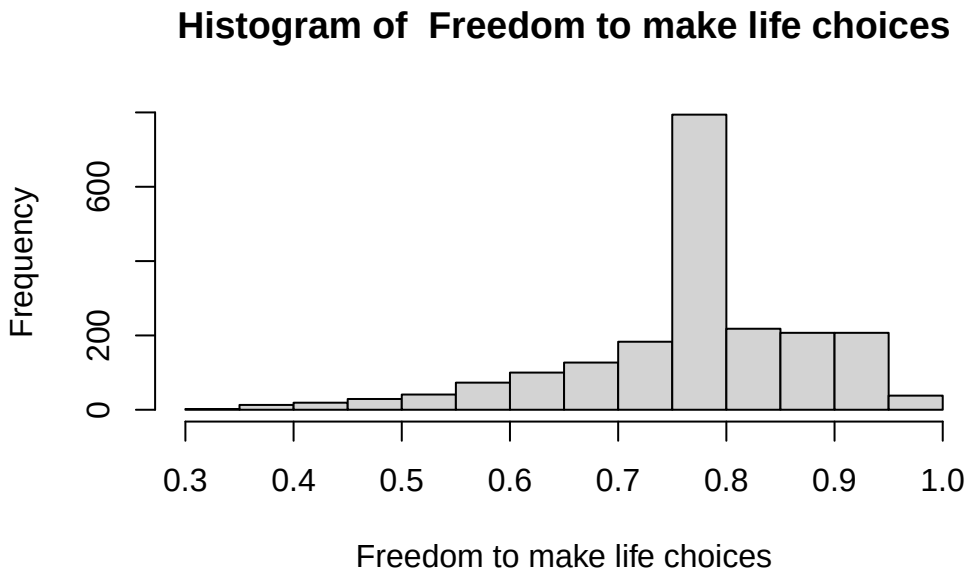

Histogram of e_peaveduc
Frequency
0
200
600
2
4
6
8
10
12
14
e_peaveduc

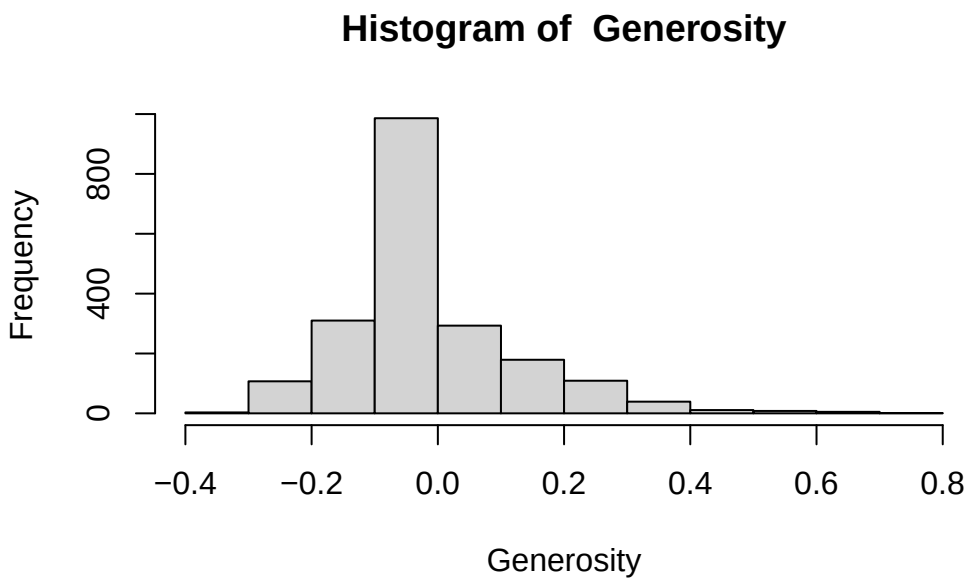

Histogram of v2x_corr
Frequency
0
100
200
300
0.0
0.2
0.4
0.6
0.8
1.0
v2x_corr


Histogram of v2x_cspart
Frequency
0
100
300
0.0
0.2
0.4
0.6
0.8
1.0
v2x_cspart


Histogram of v2xcs_ccsi
Frequency
0
200
400
0.0
0.2
0.4
0.6
0.8
1.0
v2xcs_ccsi


Histogram of Social support
Frequency
0
200
600
0.2
0.4
0.6
0.8
1.0
Social support


Histogram of Freedom to make life choices
Frequency
0
200
600
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Freedom to make life choices




```
# Normalize the data that are skewed

install.packages("bestNormalize")
library("bestNormalize")

# # Normalize the data
# We will try this again with the mice() package too

# v2c_BN <- bestNormalize(new_total$v2csreprss, r = 1, k = 5)
# new_total$v2csreprss <- v2c_BN$x.t
#
# # Check the data to see if it is normalized
# hist(new_total$v2csreprss)
#
# colnames(new_total)
#
# # Repeat for other columns
# ecivil_BN <- bestNormalize(new_total$e_civil_war, r = 1, k = 5)
# new_total$e_civil_war <- ecivil_BN$x.t
#
# ecoup_BN <- bestNormalize(new_total$e_pt_coup, r = 1, k = 5)
# new_total$e_pt_coup <- ecoup_BN$x.t
#
# v2x_BN <- bestNormalize(new_total$v2x_partipdem, r = 1, k = 5)
```

```
# new_total$v2x_partipdem <- v2x_BN$x.t
#
# epea_BN <- bestNormalize(new_total$e_peaveduc, r = 1, k = 5)
# new_total$e_peaveduc <- epea_BN$x.t
#
# v2xcorr_BN <- bestNormalize(new_total$v2x_corr, r = 1, k = 5)
# new_total$v2x_corr <- v2x_corr$x.t
```

Code book - README.qmd - Kaori

- Updated in the data folder and this folder

Data Visualization - Ephrata and Kaori

How well does it measure the outcome you are interested in? - Kaori

We chose civil society participation as our outcome variable. We were originally going to use civil society index, but we were not able to find much information regarding the actual measures considered, so we chose to use civil society participation instead. This is because civil society index may have already considered factors we are adding to the model, while civil society participation will tell us how many people are actually participating in civil society, which is the public sphere between the private sphere and government, which can include professional organizations, charities, labor unions, and spiritual groups.

What is our hypothesis? - Kaori

Our hypothesis is that countries with similar characteristics regarding supporting/ not supporting civil society - such as the presences of war or good educational systems - will have similar civil society participation rates.

How do our variable choices relate to our hypotheses? - Kaori

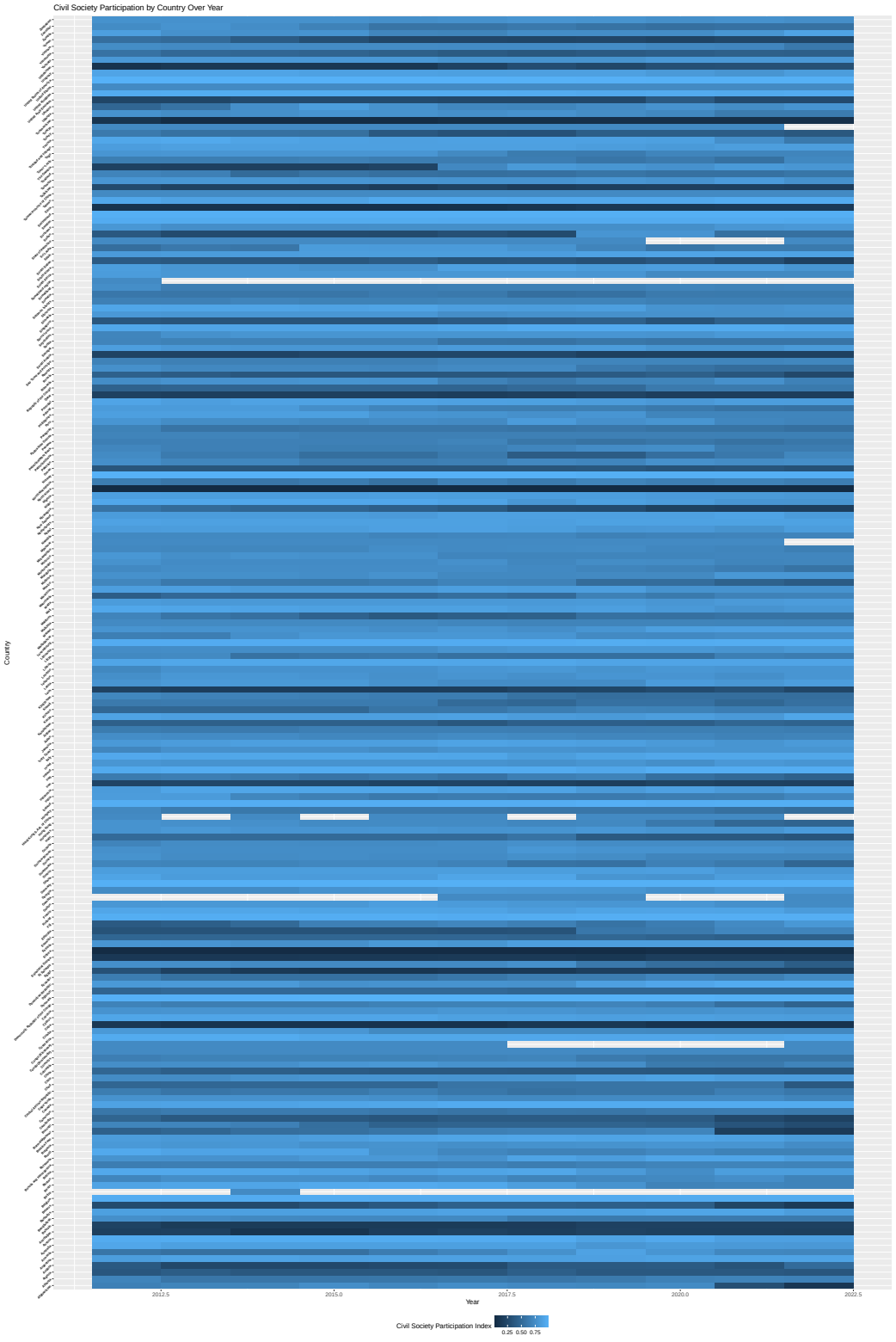
Our choices of variables reflect our understanding of civil society and the factors that make it stronger or weaker. Things that bring people together, such as generosity, and things that may bring people apart, such as corruption, we expect to have an effect on participation in civil society, which is why they are included. Civil society repression effort is included because we expect higher repression efforts to have an effect on participation in general, likely negative. This is considered in the civil society index, but will be used in preliminary analysis for better

understanding and exploration of the topic. Social support we expect to have a positive relationship with civil society participation because people will be connected with each other, and the same applies to generosity, with the opposite logic applying to presence of war and coups. Civil society index is included because we expect a stronger civil society to have a stronger participation rate, as with education because schools and teaching organizations often are a large part of civil society. The civil society index takes into account the entry/exit of civil society organizations (CSOs), the repression of civil society, and the participatory environment, which is the types and amount of CSOs available. Government corruption is included because this may dissuade people from participating in civil society or weaken civil society by underfunding it making people less likely to engage in it. Finally, participation in democracy is used because it measures the active level of citizen participation which considers suffrage, direct democracy, engagement with civil society organizations, and subnational elected bodies.

- A correlation heat map - Kaori

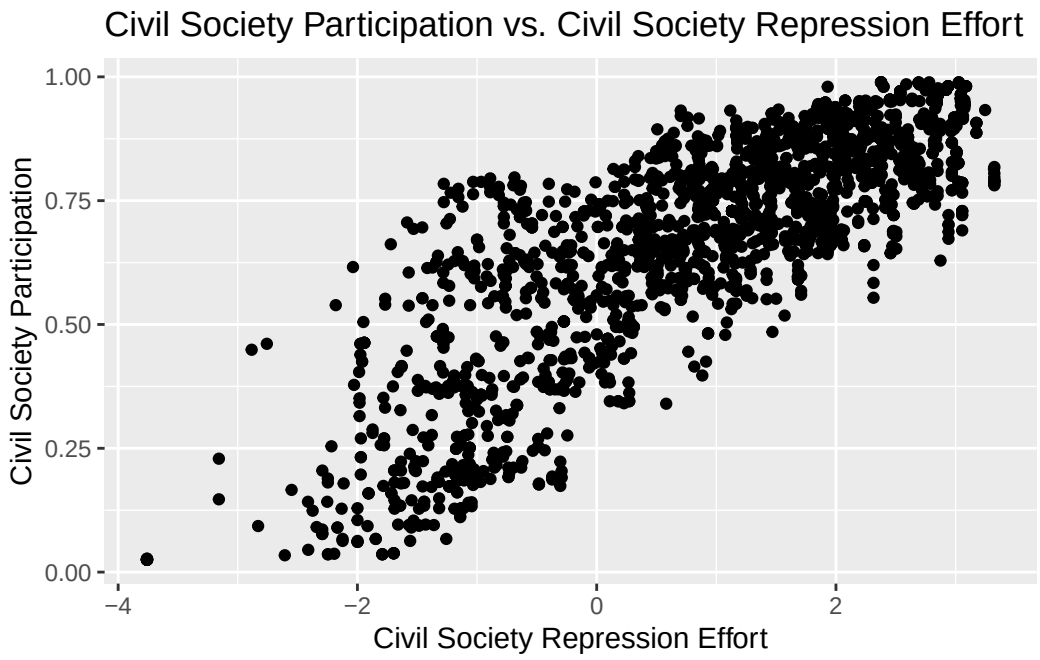
```
# changes names to shorter and easier to type/remember forms
cs <- new_total %>% rename(csrepress = v2csreprss, civil_war = e_civil_war, coup = e_pt_co
                           edu = e_peaveduc, corr = v2x_corr, cspart = v2x_cspart,
                           cs_index = v2xcs_ccsi, social_support = 'Social support',
                           choices = 'Freedom to make life choices', gen = Generosity)

cs <- cs %>% filter(year > 2011) # updates to match year for both datasets
# the full measures for happiness start in 2012, so all years before then are
# filtered out for the sake of analysis

saveRDS(cs, file = "civil_society")
```

```
# plot of civil society participation by country and year
gg <- ggplot(data = cs, aes(x = year, y = country_name, fill = cspart)) +
  geom_tile() +
  labs(title = "Civil Society Participation by Country Over Year", x = 'Year',
       y = 'Country', fill = 'Civil Society Participation Index') +
  theme(axis.text.y = element_text(face="bold", color="black",  size=5, angle = 45, hjust
  theme(legend.position = 'bottom')
  #options(repr.plot.width = 15, repr.plot.height = 50)

gg
```

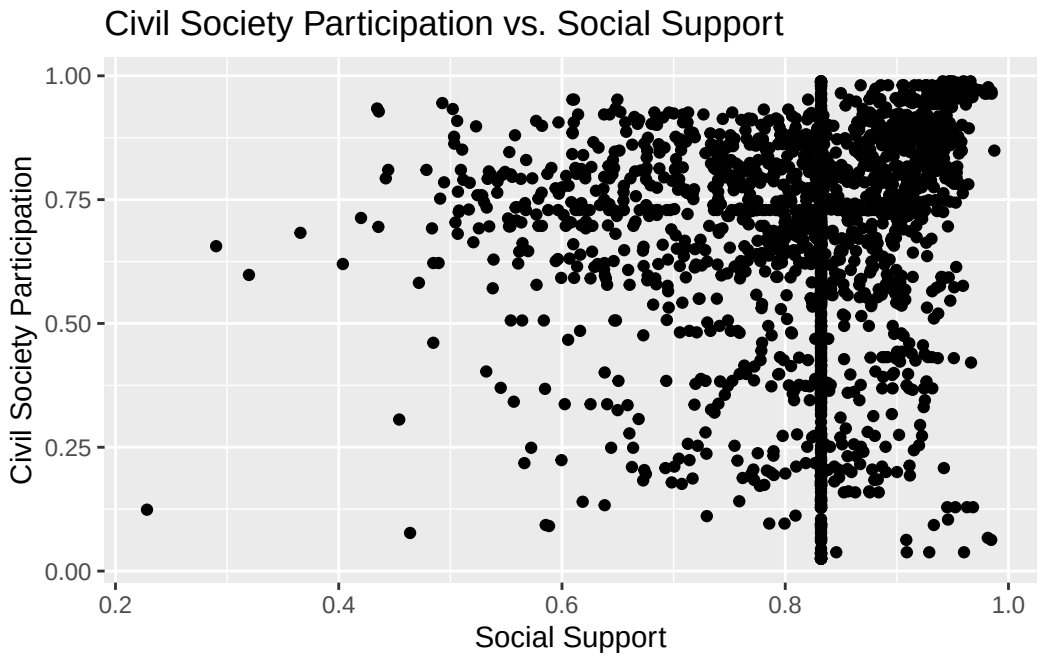

Civil Society Participation by Country Over Year
Country
2012.5
2015.0
2017.5
2020.0
2022.5
Year
Civil Society Participation Index
0.25
0.50
0.75

Visualization group 2: - Scatter plots civil society participation vs. each single predictor - Ephrata - Line graphs showing changes over time for each predictor vs. outcome - Ephrata - QQ Plot - Ephrata

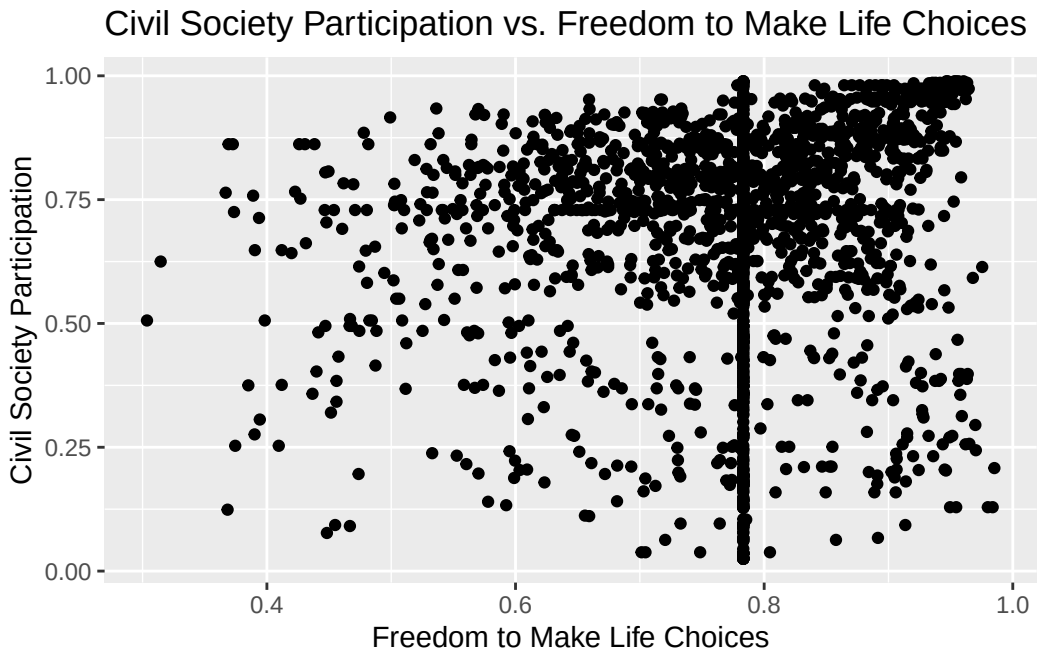
```
# Scatter plot of civil society participation vs. civil society repression effort
ggplot(cs, aes(x = csrepress, y = cspart)) +
  geom_point() +
  labs(x = "Civil Society Repression Effort", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Civil Society Repression Effort")
```




```
# Scatter plot of civil society participation vs. social support
ggplot(cs, aes(x = social_support, y = cspart)) +
  geom_point() +
  labs(x = "Social Support", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Social Support")
```

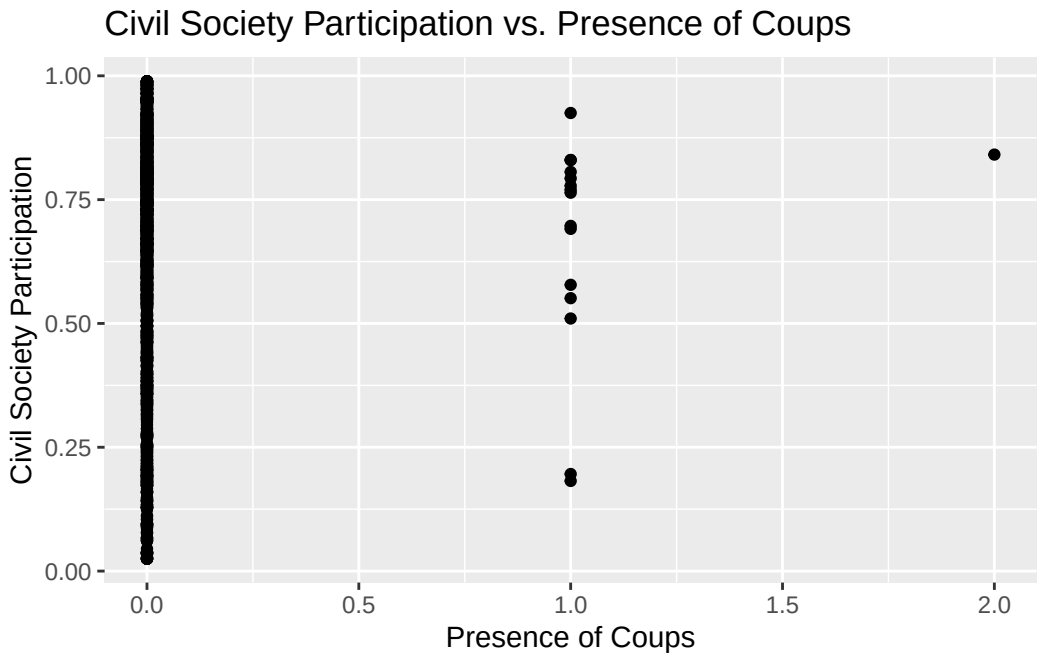



```
# Scatter plot of civil society participation vs. freedom to make life choices
ggplot(cs, aes(x = choices, y = cspart)) +
  geom_point() +
  labs(x = "Freedom to Make Life Choices", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Freedom to Make Life Choices")
```

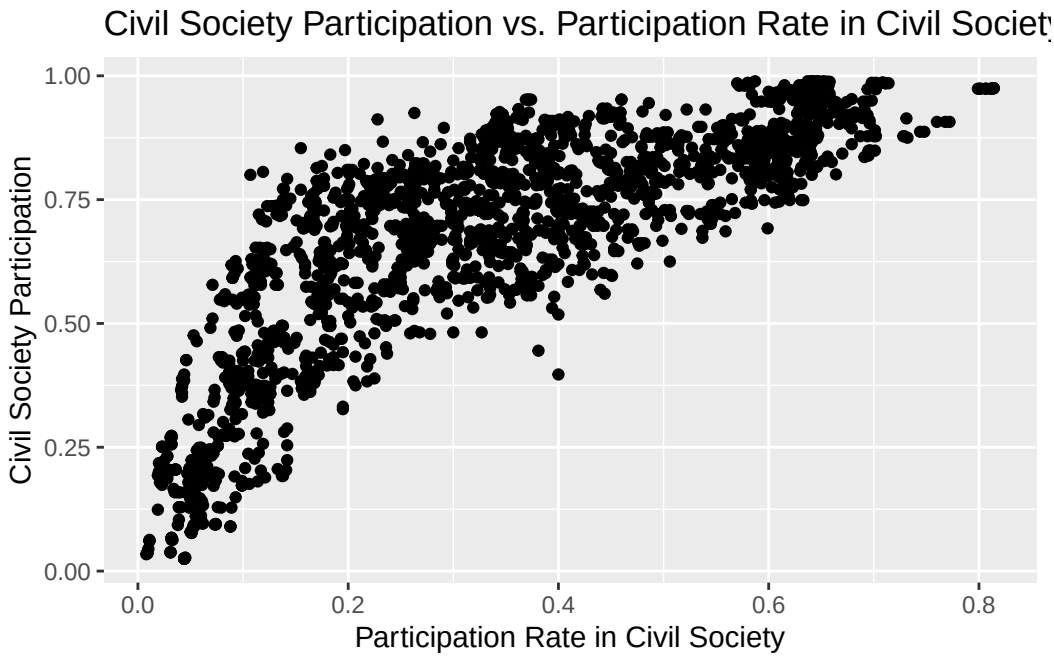



```
# Scatter plot of civil society participation vs. presence of war
ggplot(cs, aes(x = civil_war, y = cspart)) +
  geom_point() +
  labs(x = "Presence of War", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Presence of War")
```

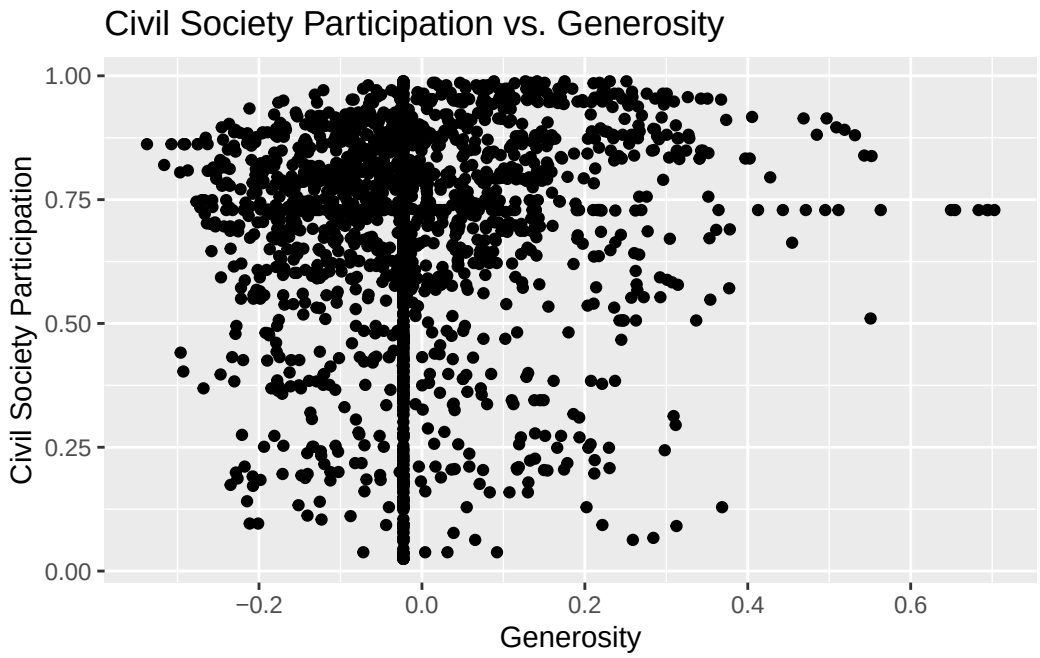

```
# Scatter plot of civil society participation vs. presence of coups
ggplot(cs, aes(x = coup, y = cspart)) +
  geom_point() +
  labs(x = "Presence of Coups", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Presence of Coups")
```



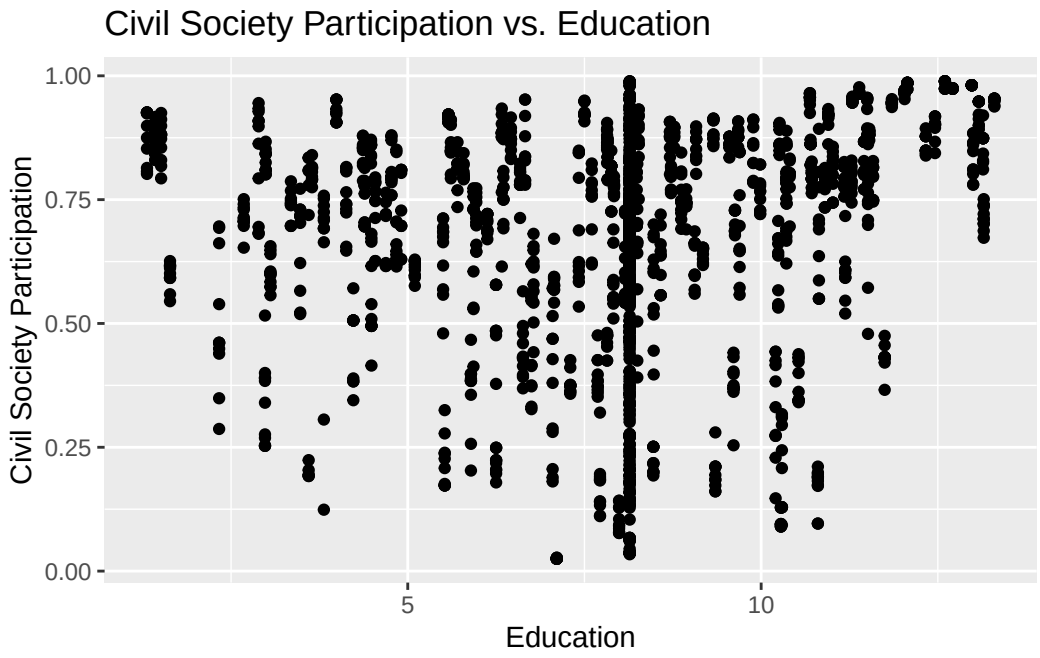

```
# Scatter plot of civil society participation vs. participation rate in civil society
ggplot(cs, aes(x = v2x_partipdem, y = cspart)) +
  geom_point() +
  labs(x = "Participation Rate in Civil Society", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Participation Rate in Civil Society")
```



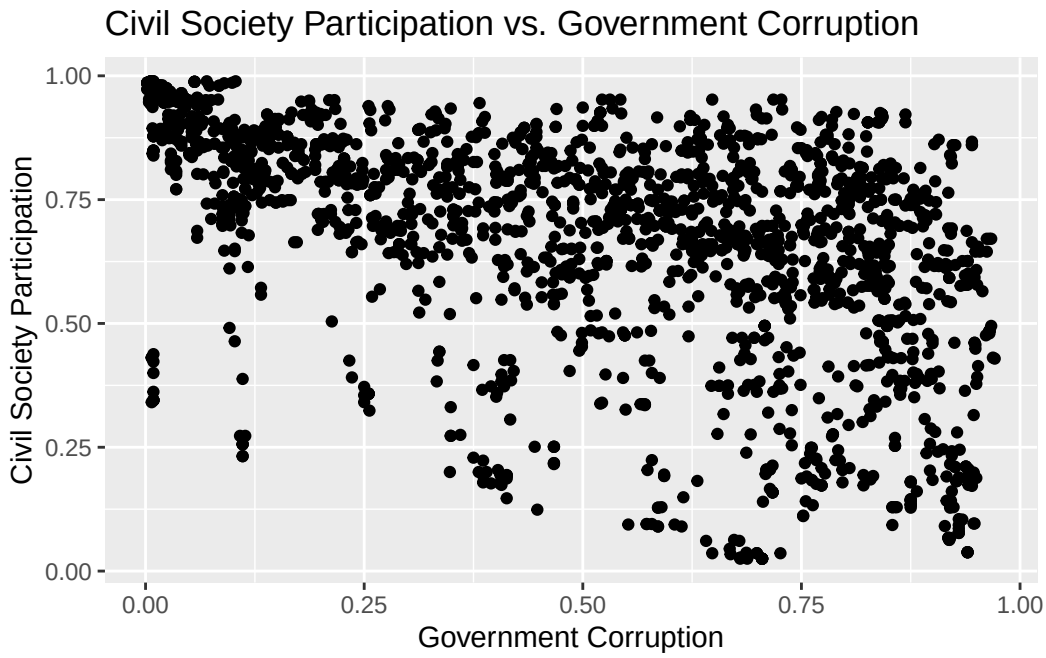

```
# Scatter plot of civil society participation vs. generosity
ggplot(cs, aes(x = gen, y = cspart)) +
  geom_point() +
  labs(x = "Generosity", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Generosity")
```



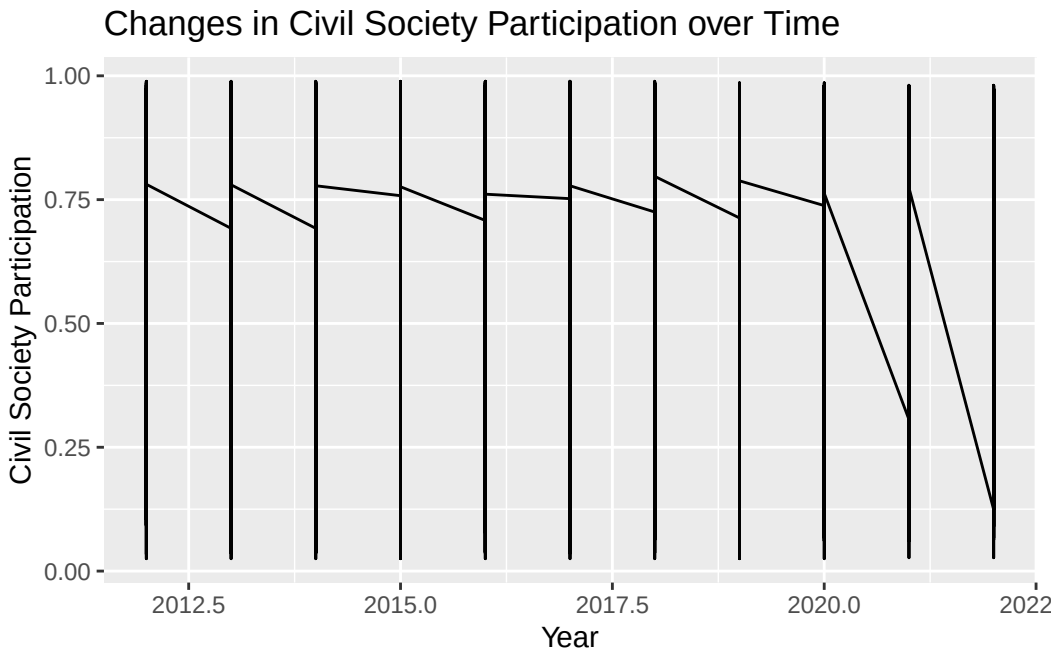

```
# Scatter plot of civil society participation vs. education
ggplot(cs, aes(x = edu, y = cspart)) +
  geom_point() +
  labs(x = "Education", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Education")
```



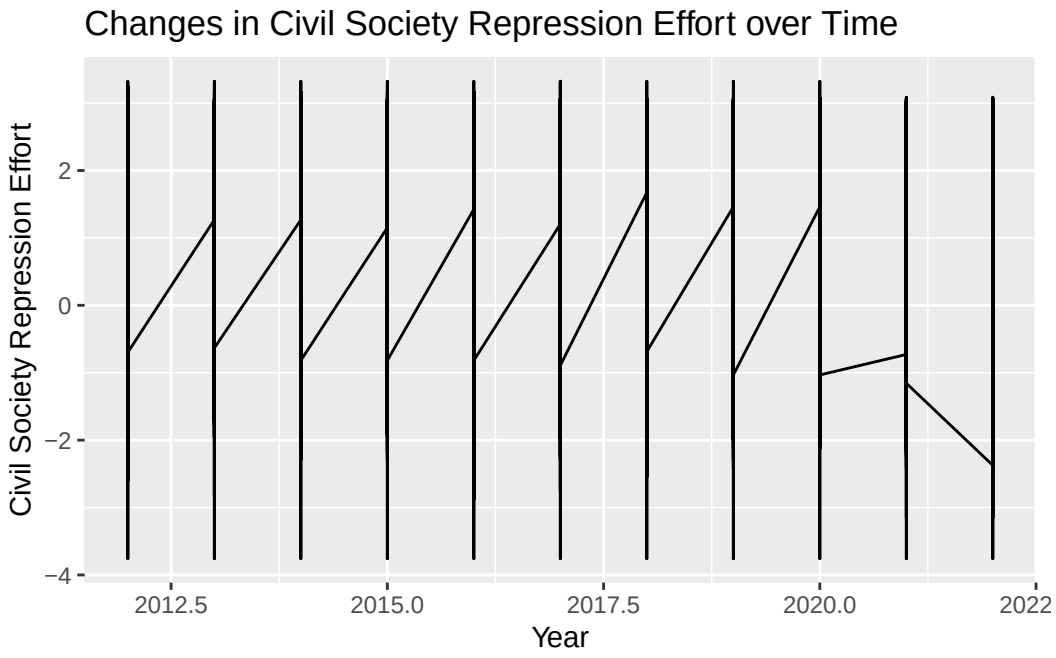

```
# Scatter plot of civil society participation vs. government corruption
ggplot(cs, aes(x = corr, y = cspart)) +
  geom_point() +
  labs(x = "Government Corruption", y = "Civil Society Participation") +
  ggtitle("Civil Society Participation vs. Government Corruption")
```



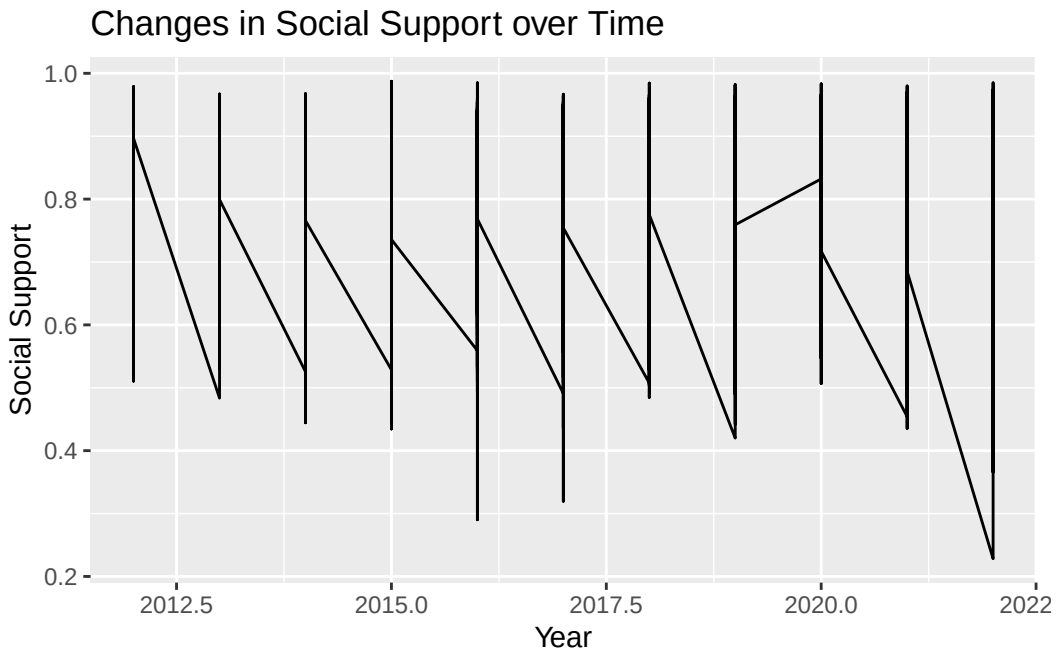

```
## the line graphs look kinda weird. is there a way to make them better for time
## series or large amounts of


# Line graph of civil society participation over time
ggplot(cs, aes(x = year, y = cspart)) +
  geom_line() +
  labs(x = "Year", y = "Civil Society Participation") +
  ggtitle("Changes in Civil Society Participation over Time")
```



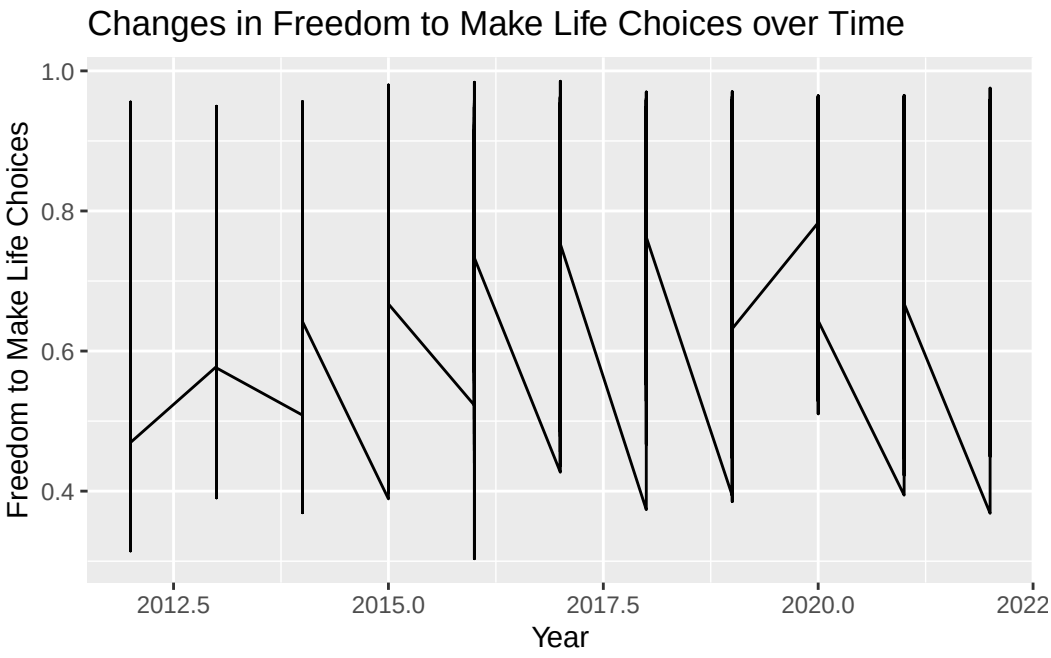

```
# Line graph of civil society repression effort over time
ggplot(cs, aes(x = year, y = csrepress)) +
  geom_line() +
  labs(x = "Year", y = "Civil Society Repression Effort") +
  ggtitle("Changes in Civil Society Repression Effort over Time")
```



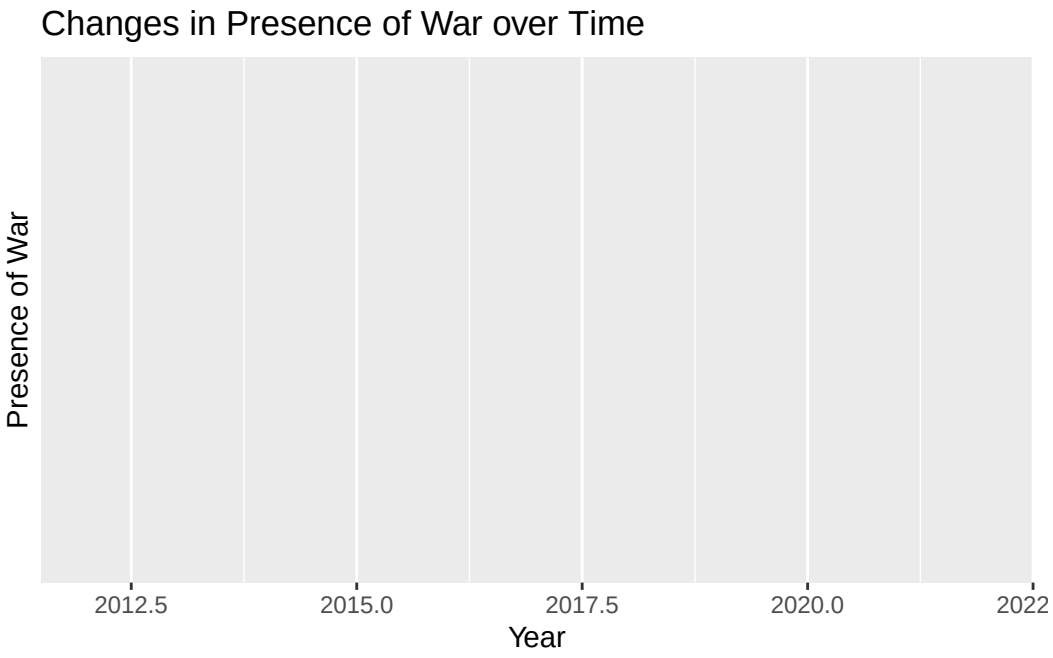

```
# Line graph of social support over time
ggplot(cs, aes(x = year, y = social_support)) +
  geom_line() +
  labs(x = "Year", y = "Social Support") +
  ggtitle("Changes in Social Support over Time")
```



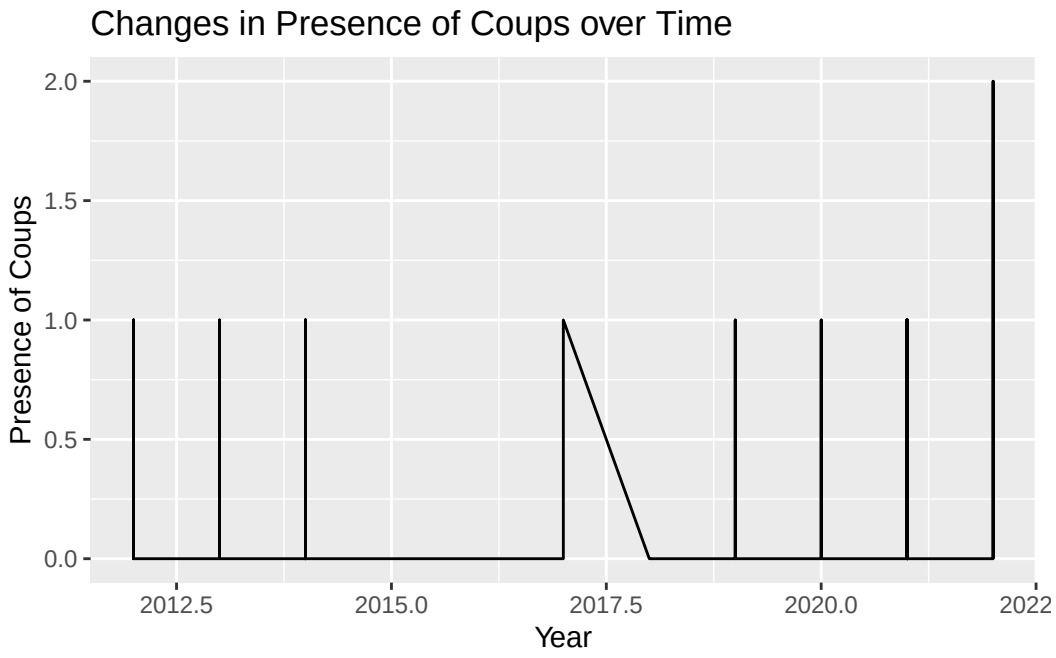

```
# Line graph of freedom to make life choices over time
ggplot(cs, aes(x = year, y = choices)) +
  geom_line() +
  labs(x = "Year", y = "Freedom to Make Life Choices") +
  ggtitle("Changes in Freedom to Make Life Choices over Time")
```



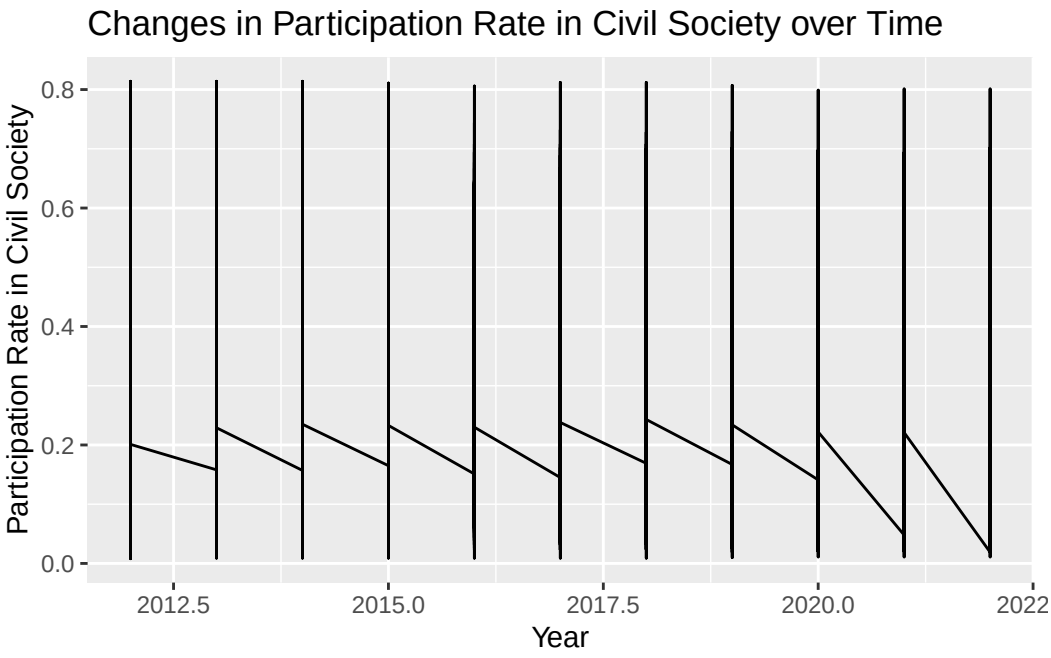

```
# Line graph of presence of war over time
ggplot(cs, aes(x = year, y = civil_war)) +
  geom_line() +
  labs(x = "Year", y = "Presence of War") +
  ggtitle("Changes in Presence of War over Time")
```



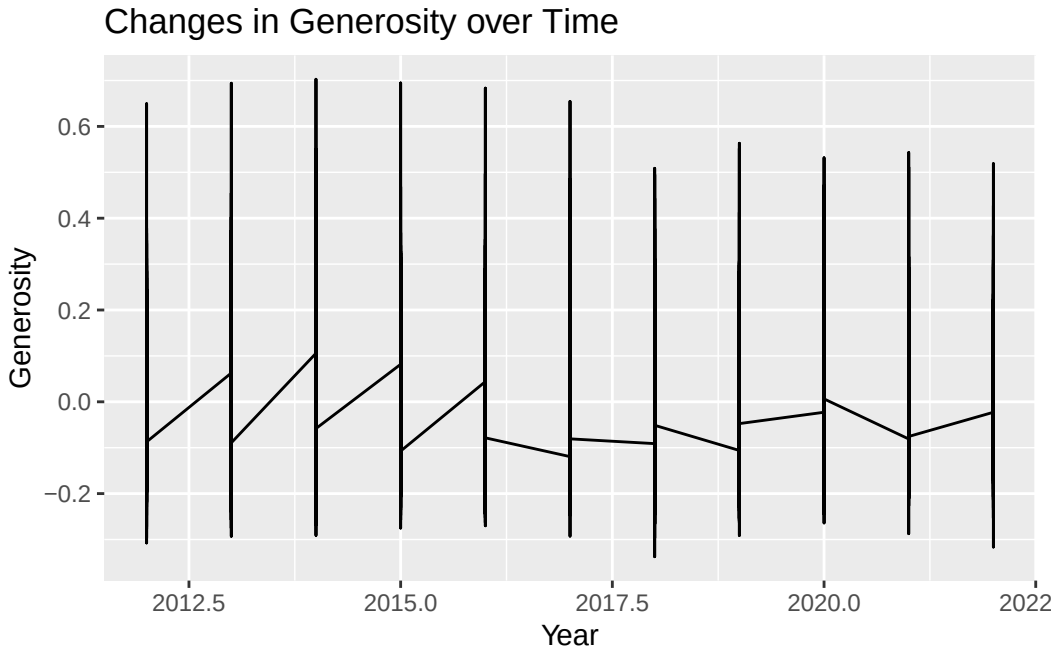

```
# Line graph of presence of coups over time
ggplot(cs, aes(x = year, y = coup)) +
  geom_line() +
  labs(x = "Year", y = "Presence of Coups") +
  ggtitle("Changes in Presence of Coups over Time")
```



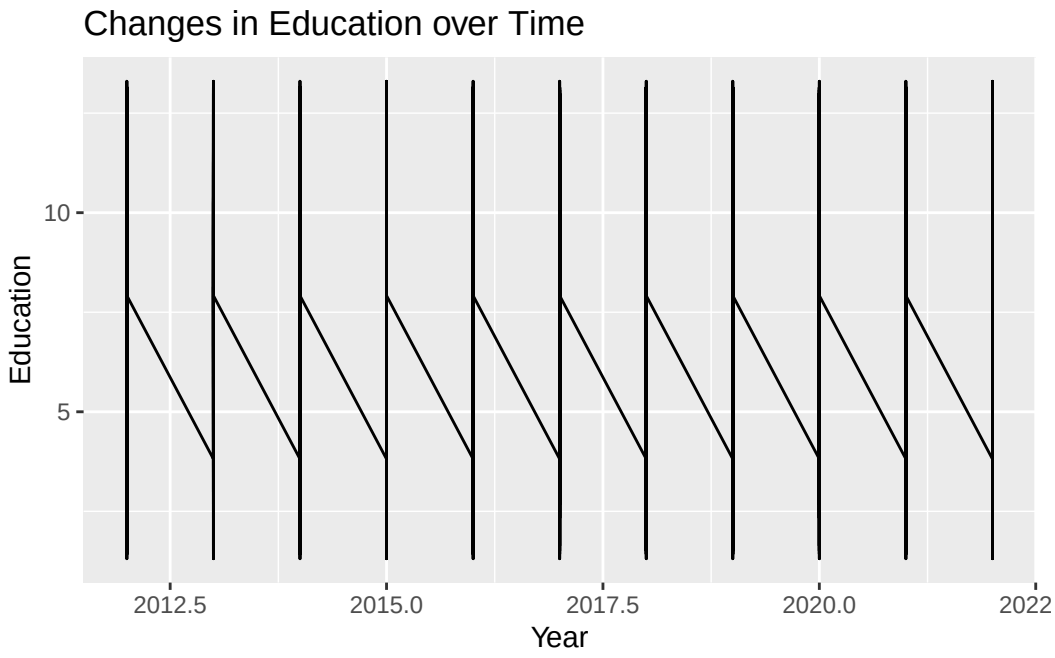

```
# Line graph of participation rate in civil society over time
ggplot(cs, aes(x = year, y = v2x_partipdem)) +
  geom_line() +
  labs(x = "Year", y = "Participation Rate in Civil Society") +
  ggtitle("Changes in Participation Rate in Civil Society over Time")
```



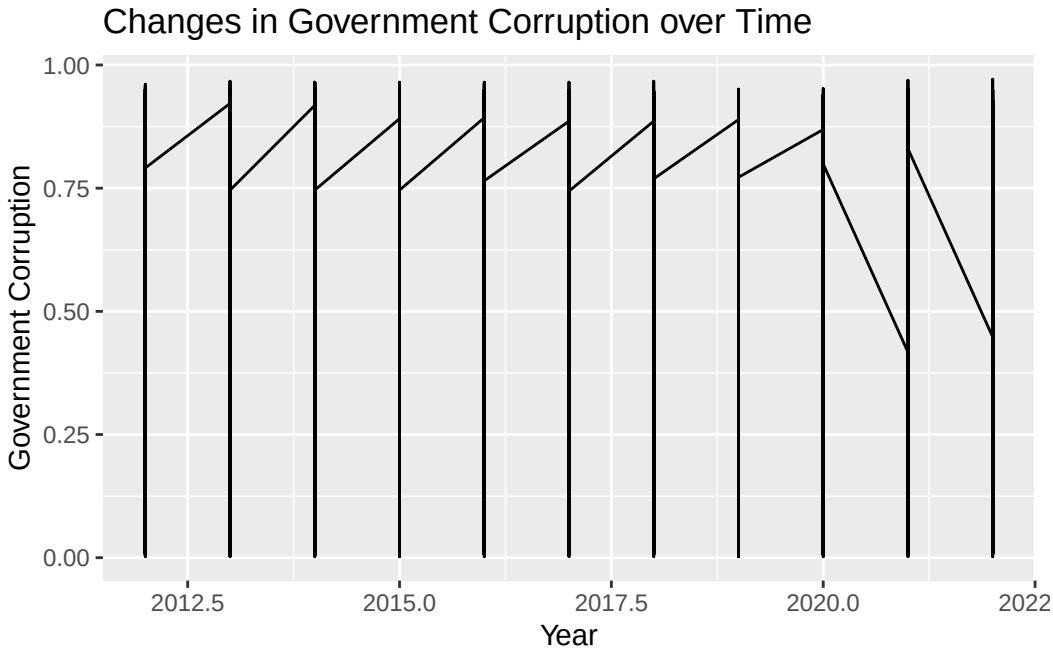

```
# Line graph of generosity over time
ggplot(cs, aes(x = year, y = gen)) +
  geom_line() +
  labs(x = "Year", y = "Generosity") +
  ggtitle("Changes in Generosity over Time")
```




```
# Line graph of education over time
ggplot(cs, aes(x = year, y = edu)) +
  geom_line() +
  labs(x = "Year", y = "Education") +
  ggtitle("Changes in Education over Time")
```




```
# Line graph of government corruption over time
ggplot(cs, aes(x = year, y = corr)) +
  geom_line() +
  labs(x = "Year", y = "Government Corruption") +
  ggtitle("Changes in Government Corruption over Time")
```




```r
# QQ plot of civil society participation
qqnorm(cs$cspart, main = "QQ Plot of Civil Society Participation")
qqline(cs$cspart)
```