Final Report

Group 4: Kaori Hirano, Alicia Nguyen, James Xia

2023-08-01

Introduction and Data

Topic & Research Question

We are interested in how the civil society participation and societies are mutually affected: Our research question is determining the most essential factors in predicting civil society participation. We hypothesized that characteristics regarding social cohesion–such as region, presence of war, social support, and civil society index–will have large influences on civil society participation rates. The World Happiness Index and V-Dem datasets were chosen because want to understand civil society better by understanding what affects it from a variety of perspectives. The World Happiness Index measures social phenomenons in a quantitative manner, and the V-Dem provides a variety of indexes measuring the stability of a society from various perspectives. 2019 is the most recent data set we have that is not reflecting social changes due to COVID-19, which is why we chose 2019 specifically over another year. This might be more representative of the social conditions before the pandemic year.

Based on Jens Steffek and Maria Paola Ferretti's publication of "Accountability" or "Good Decisions"? The Competing Goals of Civil Society Participation in International Governance" on 07 Jan 2009, "(civil society participation) may either enhance the democratic accountability of intergovernmental organisations and regimes, or the epistemic quality of rules and decisions made within them." They pointed out the direct cause and effect relationship between civil society participation and the shape of societies. To examine this, we chose to look at the question from both a social and political lens.

Data Information

Data is attributed to the Varieties of Democracy (V-Dem) Project which provides a multi-dimensional perspective on democracy beyond elections by considering measures including electoral, participatory, egalitarian, deliberative, and liberal principles in data collection and the Sustainable Development Solutions Network which aggregates data from the Gallup World

Poll on happiness, GDP per capita, social support, healthy life expectancy, freedom, generosity, and corruption. Further information on the measures used in V-Dem specifically is found in the V-Dem codebook. Measures for the happiness data were found in the 2023 statistical appendix. These datasets are collected yearly, with V-Dem having data dating back hundreds of years (since 1789) while the happiness index dataset has been collected since 2019. We specifically used data from 2019. The V-Dem group collects the data themselves. The world happiness index has information from a variety of sources, including gallup. Many elements of the datasets are created from multiple or different variables, such as civil society index which is defined by several factors, including number of civil society organizations and the length such civil society organizations exist. Explanations for how data categories are created and calculated are clearly laid out in the respective codebooks or statistical appendixes. Each row represents a country in the year of 2019, with the relevant data associated with it from the V-Dem and happiness index datasets.

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. Additionally, 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. Additionally, we chose three more social variables from the happiness dataset that we thought would relate to civil society participation: life ladder, longevity, and GDP per capita. Happiness is measured by the life ladder, which is a scale from 0-10 which represents the worst to best possible life and takes the average of all scores in a country. Longevity is the life expectancy at birth for a person in this year. Finally, is GDP, which is log GDP per capita, GDP calculated at constant 22018 international dollar prices. These three were added because we suspect that a society with longer life spans, higher happiness overall, and more money would be associated with people having more ability to spend time participating in civil society.

Data Cleaning and Wrangling

As for data wrangling, no variables were significantly altered from the original dataset. The only variable change that was made is changing the regions from numbers (1-10) to the region names. NAs were dealt with by replacing the NA with the median of that column in 2019. This is not the best method of dealing with NAs but we reduced the window we drew the median from to just 2019 instead of the entire dataset to reduce the possibility of skewing the data by incorporating information that may not fit into the context of the rest of the 2019 data. Further, NA values were replaced with the median over the mean to ensure we do not skew the data considering that most of the variables were not normally distributed. Variables were scaled when used in the Lasso and ridge regression models. As seen in the summary statistics, there were wide distributions between predictors, making scaling necessary for best accuracy. Note that scaling is not necessary in tree models, which is why it was not scaled there. Scaling changes the interpretation of the results, so rather than one unit of the variable having the listed effect on the outcome it becomes one standard deviation of the variable having the listed effect on the outcome.

All variables except civil war and coup are going to be used in the models. Civil war and coup were dropped because there were no instances of civil war or coups in 2019. Because of this, including these variables did not make sense. The reasoning for including all other variables is because we believed these variables were likely to have an effect on civil society participation rates. None of the models we chose had limitations in variables allowed, such as numerical only, so we were able to successfully add all of them to the models. However, to make our models be the most effective, we did have to deal with NA values.

When creating regression models, we had to specify a baseline level of region because it is a factor. The median and mode values of the outcome variable across regions seem to be approximately within the same vicinity. This means that as long as we choose one of the regions not with abnormal civil society participation as the reference, our analysis will not be affected. Similarly, the reference level is not incredibly important as we do not care about a specific region, we simply want to know if regional differences do exist which would likely be present regardless of the baseline chosen. Because the graphs show that Sub Saharan Africa is largely in the middle of the regions, we chose it as our reference level.

```r
wh_2023 <- read_excel("/cloud/project/data/wh_2023.xls")
load("/cloud/project/data/vdemdata-master/data/vdem.RData")
## load the original datasets

need_vdem=c("country_name","year","v2csreprss",'e_civil_war',"e_pt_coup",
            "v2x_partipdem","e_peaveduc","v2x_corr","v2x_cspart","v2xcs_ccsi",
            'e_regionpol')
vdem_use=vdem[,need_vdem]
need_wh=c("Country name","year","Social support","Freedom to make life choices",
          "Generosity","Life Ladder","Healthy life expectancy at birth",
          "Log GDP per capita")
wh_use=wh_2023[,need_wh]
## only select the portions where we need to need for our project

colnames(wh_use)[1] <- "country_name"
## change col name for eaier merge

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

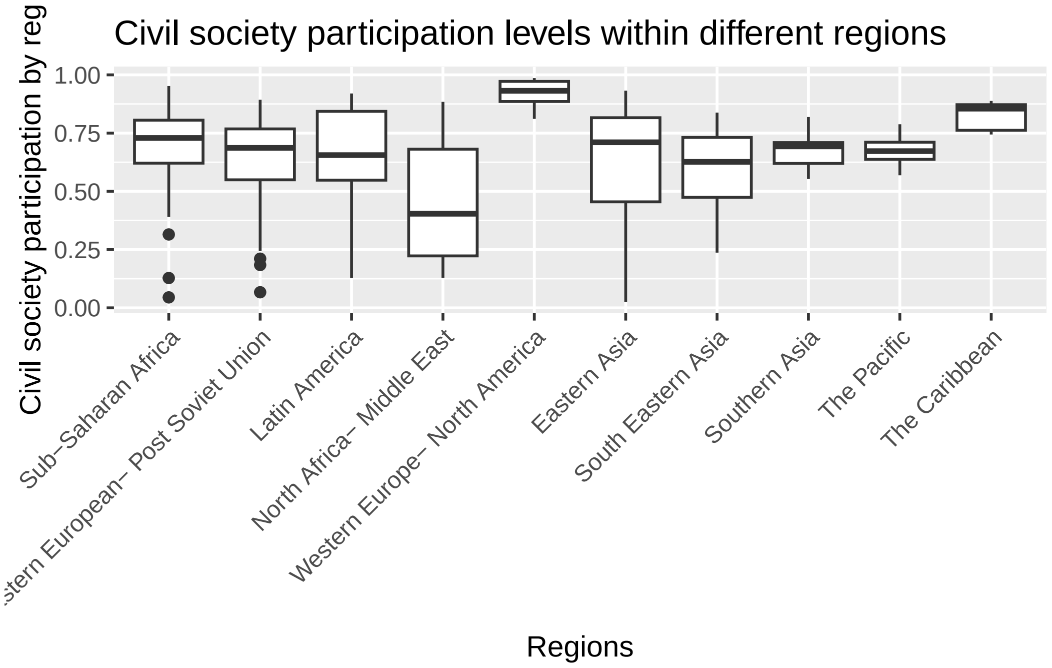
```
# Select the time range from 2019 only
total_2019 <- total[total$year == 2019,]

#replace missing values in each numeric column with median value of column
total_2019 <- total_2019 %>%
  mutate(across(where(is.numeric),~replace_na(.,median(.,na.rm=TRUE))))

# removing coup and civil war because it was all NA or 0 for 2019
total_2019 <- total_2019 %>% select(-one_of("civil_war", "coup"))

# changes names to shorter and easier to type/remember forms
cs <- total_2019 %>% rename(csrepress = v2csreprss, civil_war = e_civil_war,
                            coup = e_pt_coup,
                            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,
                            region = e_regionpol,
                            happ = "Life Ladder",
                            lifee = "Healthy life expectancy at birth",
                            gdp = "Log GDP per capita")
```

```
# sets as factor
cs$region <- as.factor(cs$region)

# We choose Sub Saharan Africa as the reference level
# We are interested in how all regions differ from one another
# We are not focused on only one specific region so we did not intentionally
# choose Sub Saharan Africa
cs$region <- relevel(cs$region, ref = 4)

# creates new data frame for tree plotting with regions as numbers still
cs_num <- cs

# recodes region to be the region it represents rather than a number value/code
levels(cs$region) <- c('SubSaharanAfrica',
                    'EasternEurope_PostSovietUnion',
                    'LatinAmerica',
                    'NorthAfrica_MiddleEast',
                    'WesternEurope_NorthAmerica',
                    'EasternAsia',
                    'SouthEasternAsia',
                    'SouthernAsia',
                    'ThePacific',
                    'TheCarribean')
```

```
# Visualize all the variables to see whether we need to scale the variables
# Because of the scale differences, we do need to scale variables

# Visualize the outcome variable to see how it is distributed
summary(cs$cspart)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0250  0.5760  0.7290  0.6740  0.8355  0.9860
```

```
# A box plot to show the relationship between region and
# Civil Society Participation

csp <- ggplot(cs, aes(x=as.factor(region),
               y = cspart,
               group = as.factor(region))) +
  geom_boxplot() +
  labs(x = "Regions",
       y = "Civil society participation by region",
       title = "Civil society participation levels within different regions") +
  # Rotate x-axis labels by 90 degrees to the left
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none") +
  scale_x_discrete(labels = c("Sub-Saharan Africa",
                              "Eastern European- Post Soviet Union",
                              "Latin America",
                              "North Africa- Middle East",
                              "Western Europe- North America",
                              "Eastern Asia",
                              "South Eastern Asia",
                              "Southern Asia",
                              "The Pacific",
                              "The Caribbean"))
csp
```

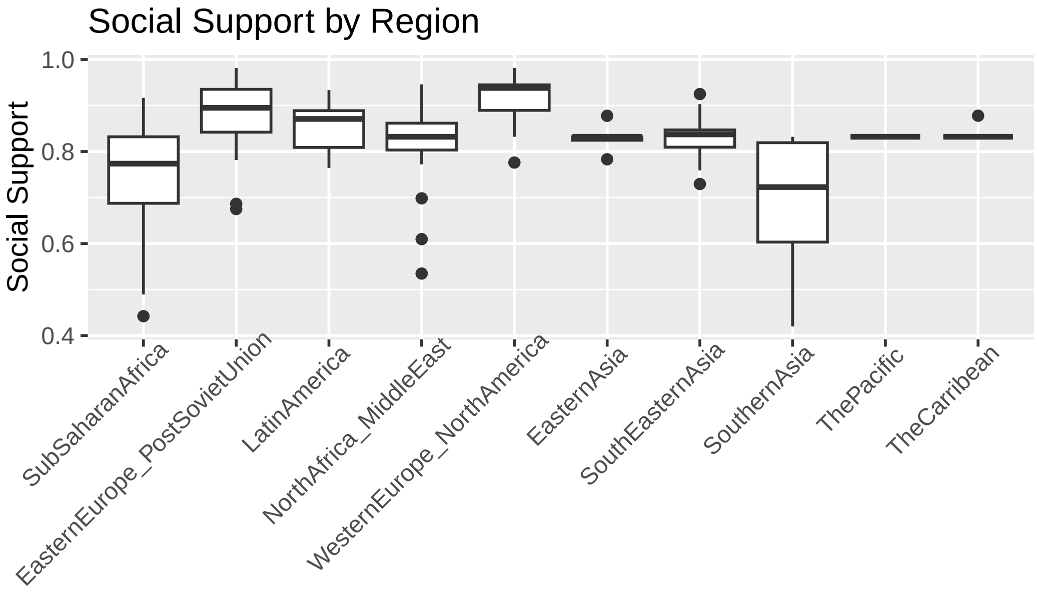


Note: This histogram is to show that the reference level for the Region variable, Sub-Saharan Africa is along the same median level for Civil Society Participation variable as the majority of other regions. This means that when we interpret the coefficient of the different Regions, the metric/coefficient is reflecting how different certain regions are, in terms of Civil Society Participation, compared to Sub-Saharan Africa and similar regions.

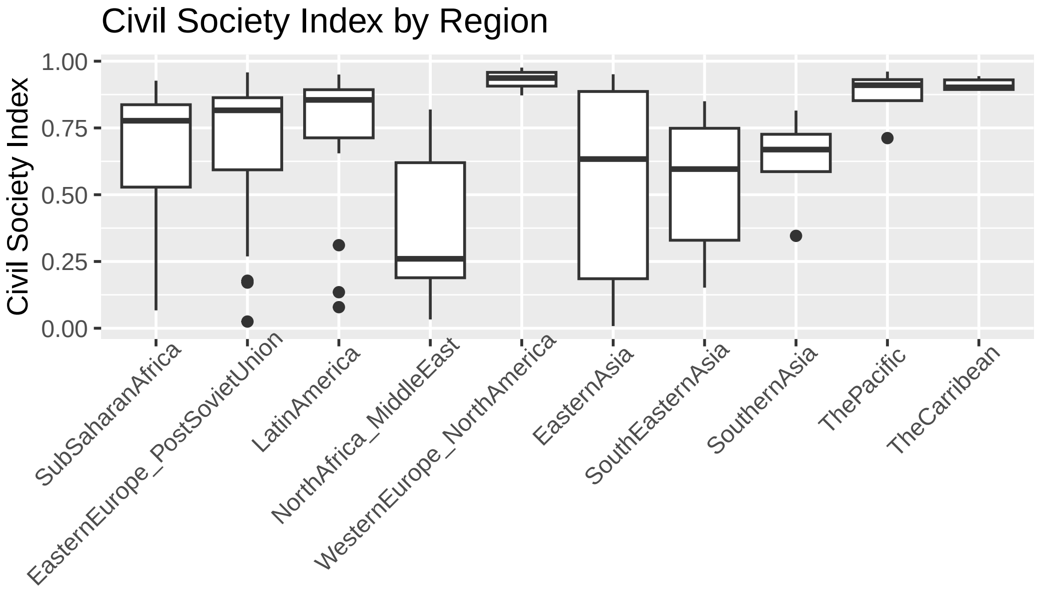
Methodology

We also plotted the distribution of different important social variables in our hypothesis against regions to see whether there are any strong correlation between our predictors and regions. This helped us visualize any effect of region on the outcome and the effect of various predictors on region. After the visualization, it is hard to point out a consistent and clear trend among the regions. There are no regions that score consistently high or low across all three of these chosen social participation related variables.

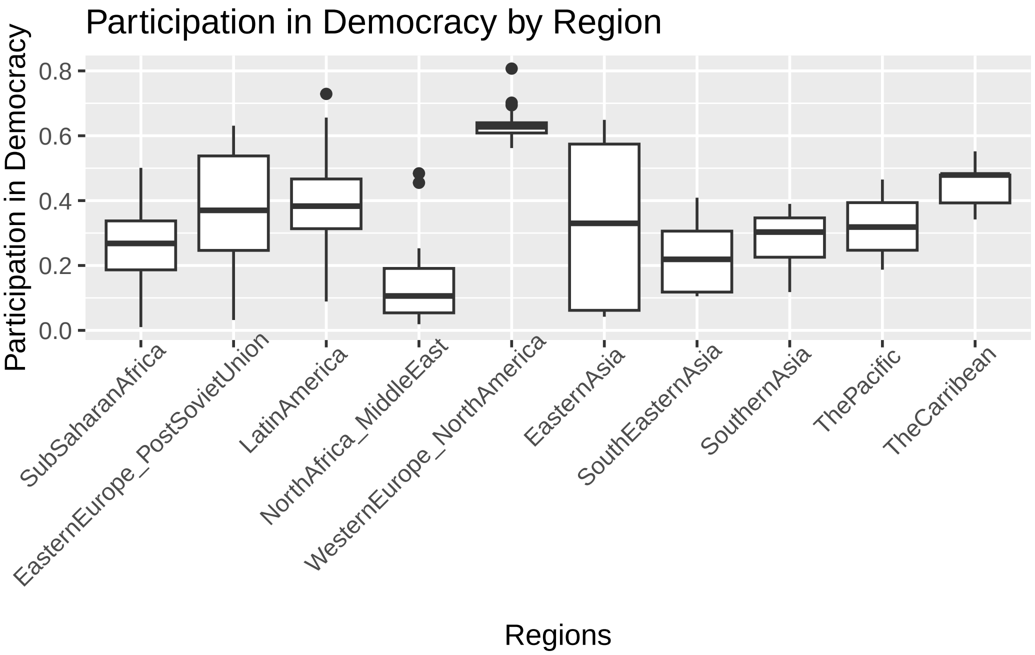
```
# A box plot to show the relationship between region and social support
social <- ggplot(cs, aes(x=as.factor(region),
               y = social_support,
               group = as.factor(region))) +
  geom_boxplot() +
  labs(x = "",
       y = "Social Support",
       color = "Region",
       title = 'Social Support by Region') +
  # Rotate x-axis labels by 90 degrees to the left
  theme(axis.text.x = element_text(angle = 45, hjust = 0.9),
        legend.position = "none")
social
```



```
# A box plot to show the relationship between region and Civil Society Index
csi <- ggplot(cs, aes(x=as.factor(region),
               y = cs_index,
               group = as.factor(region))) +
  geom_boxplot() +
  labs(x = "",
       y = "Civil Society Index",
       color = "Regions",
       title = 'Civil Society Index by Region') +
  # Rotate x-axis labels by 90 degrees to the left
  theme(axis.text.x = element_text(angle = 45, hjust = 0.9),
        legend.position = "none")
csi
```

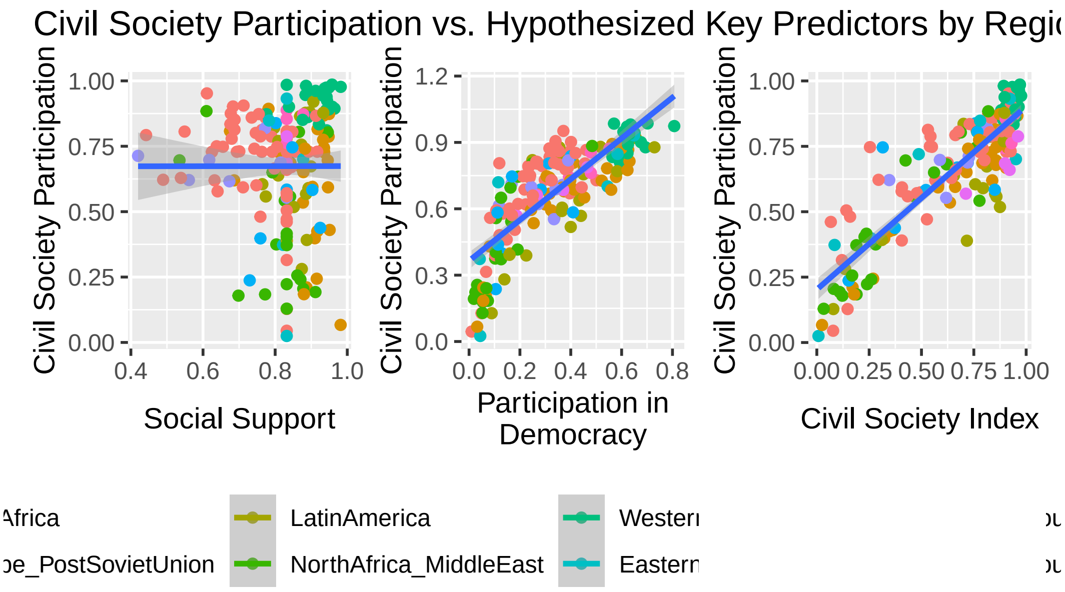


```
# A box plot to show the relationship between region
# and participation in democracy
democracy <- ggplot(cs, aes(x = as.factor(region),
               y = v2x_partipdem,
               group = as.factor(region))) +
  geom_boxplot() +
  labs(x = "Regions",
       y = "Participation in Democracy",
       color = "Regions",
       title= 'Participation in Democracy by Region') +
  # Rotate x-axis labels by 90 degrees to the left
  theme(axis.text.x = element_text(angle = 45, hjust = 0.9),
        legend.position = "none")
democracy
```



We also plotted the expected most important variables in our hypothesis against the outcome to see whether there is a complex relationship or a linear one between them. The method *lm* for the best fit line was not specified as we wanted to understand the type of relationship that may be present. We are also interested in how strong the relationship is. There is a strong and more linear relationship between civil society participation and civil society index than between the outcome variables and the other two chosen variables. The other two relationships are not as linear and more complicated, as visualized below. This information may help us interpret our models later on.

```
# A scatter plot to show the relationship between civil society participation
# and social support by region

# Reformat the legend
# Custom labels for color legend with text wrapping
custom_labels <-            c("Sub-Saharan Africa",
                              "Eastern European- Post Soviet Union",
                              "Latin America",
                              "North Africa- Middle East",
                              "Western Europe- North America",
                              "Eastern Asia",
                              "South Eastern Asia",
                              "Southern Asia",
                              "The Pacific",
                              "The Caribbean")

viz1 <- ggplot(cs, aes(x=social_support,
               y = cspart,
               color = region)) +
  geom_jitter() +
  geom_smooth(method = 'lm', aes(group = 1)) +
  labs(x = "Social Support",
       y = "Civil Society Participation") +
  theme(legend.position = "none")

# A scatterplot to show the relationship between civil society participation
# and participation in democracy by region
viz2 <- ggplot(cs, aes(x=v2x_partipdem,
               y = cspart,
               color = region)) +
  geom_point() +
  geom_smooth(method = 'lm', aes(group = 1)) +
  labs(x = str_wrap("Participation in Democracy",25),
       y = "Civil Society Participation") +
  theme(legend.position = "bottom")

# A scatterplot to show the relationship between civil society participation
# and civil society index by region
viz3 <- ggplot(cs, aes(x=cs_index,
               y = cspart,
               color = region)) +
  geom_point() +
  geom_smooth(method = 'lm', aes(group = 1)) +
  labs(x = "Civil Society Index",
       y = "Civil Society Participation",
       color = "Region") +
  theme(legend.position = "none") +
  scale_color_discrete(labels = c("Sub-Saharan Africa",
                              str_wrap("Eastern European- Post Soviet Union", 20),
                              "Latin America",
                              str_wrap("North Africa- Middle East", 13),
                              str_wrap("Western Europe- North America", 15),
                              "Eastern Asia",
                              "South Eastern Asia",
                              "Southern Asia",
                              "The Pacific",
                              "The Caribbean"))

## Note that when rendered these three are too squished to be read.
## might be better to just do them all on their own?
viz1 + viz2 + viz3 +
  plot_annotation(title =
  "Civil Society Participation vs. Hypothesized Key Predictors by Region",
  caption = "Blue lines represent the best fit line for the respective plots.")
```



Blue lines represent the best fit line for the respective plots.

Model Choice

Our research question regards which predictors are the most important in predicting civil society participation. We further edited this question to focus on 2019, specifically, to make our analysis more feasible.

Our outcome variable is a numerical variable. We chose to use regression to identify the most important predictors of civil society participation in 2019. Lasso will eliminate non-influential variables, while ridge will shrink them to a smaller value, both of which minimize the presence of non-influential predictors. We will do both models to test which of our chosen variables are important in predicting civil society participation. To optimize our model performance, the optimal lambda will be determined using 10-fold cross validation.

We also used tree models to further understand the most important variables. This is because, as shown in our data visualizations, there are not only linear relationships between the predictors and the outcome variable. So, doing a different type of modelling might help us shed light on the relationship between our predictors and the outcome variable civil society participation.

We initially chose a single decision tree to give a base understanding of the main predictors visually. However, because single decision trees are not the best in terms of prediction compared to other tree models, we also used a random forests model. Random forests allows us to see a relative ranking of the most important variables and their relationships to each other in our model. The hyperparameters (terminal nodes for the single decision tree and number of variables considered for random forests) will be chosen by 5-fold cross validation to minimize error.

If the topmost important variables produced by our machine learning models are consistent with our hypothesis, and the model works well when applied to the test set, this means that the model results are relatively consistent with our hypothesis. We also want to see the nature of the specific relationships–whether the relationship is negative or positive–and the relative importance of the relationship between civil society participation and each of the predictor variable to help us understand our research question.

Results

The best performing model was random forests, which had the highest R-squared value and the lowest MSE on the test set. Across all models, civil society index, participation in democracy, and region were identified as the most influential in predicting civil society participation. Other important variables identified by a non-majority include social support, life expectancy, and civil society repression. The specific results for each model and comparison between the model types are listed below.

Test Train Split

```
# 70, 30 test train split
set.seed(145)
train <- sample(c(TRUE, FALSE), nrow(cs), replace = TRUE, prob=c(.7,.3))
test <- (!train)
val <- test
```

Regression

```
# create x and y for glmnet
set.seed(129)
x <- model.matrix(cspart ~ csrepress+v2x_partipdem+edu+corr+cs_index+
                           social_support+choices+gen+region+lifee+happ+gdp,
                  data = cs)[, -1]
y <- cs$cspart
```

Ridge Regression

```
# set seed for reproducibility
set.seed(129)

# cross validation for best l
cv_r <- cv.glmnet(x[train,], y[train], alpha = 0,
                  # scale the x values
                  lambda = 10^seq(10, -2, length = 100), scale = TRUE)

# saving optimal lambda
bestlam_r <- cv_r$lambda.min

# calculating MSE
ridge_pred <- predict(cv_r, s = bestlam_r,
newx = x[test, ], scale = TRUE)
ridge_mse <- mean((ridge_pred - y[test])^2)

# fits final ridge model
ridge_mod <- glmnet(x, y, alpha = 0, lambda = bestlam_r, scale = TRUE)
```

```r
# saves coefficients
coef_r <-coef(ridge_mod)

# fits linear ridge model
library(ridge)
lin_ridge <- linearRidge(cspart ~ . -year -country_name, data = cs[train,])
pvals_ridge <- pvals(lin_ridge)
```

```r
# NEED TO ADD P VALUES TO TABLE
# prints important coefficients as a table
ridge_feature_estimate <- ridge_mod %>% tidy() %>%
  select(term, estimate)
# remove intercept
ridge_feature_estimate <- ridge_feature_estimate[-1,]
# add
ridge_feature_estimate <- ridge_feature_estimate %>% add_column()
ridge_feature_estimate_ordered <- ridge_feature_estimate %>%

  arrange(desc(abs(pval)))
```

```
Error in `arrange()`:
i In argument: `..1 = abs(pval)`.
Caused by error:
! object 'pval' not found
```

```r
print(ridge_feature_estimate_ordered)
```

```
Error in h(simpleError(msg, call)): error in evaluating the argument 'x' in selecting a metho
```

```r
  arrange(desc(abs(estimate))) # %>%
```

```
Error in eval(expr, envir, enclos): object 'estimate' not found
```

```r
  # mutate(term = c())
#print(ridge_feature_estimate_ordered)
```

```r
## WE NEED TO CHANGE THE ROUNDING
ridge_feature_estimate_ordered_table=ridge_feature_estimate_ordered%>%
  mutate(term = c('Intercept',
                  "Civil Society Index",
                  "Democracy Participation",
                     "Social Support",
                      "The Pacific",
                    "Latin America",
                    "Eastern European- Post Soviet Union",
                    'Freedom to make life choices',
                    "Gov Corruption Index",
                    "The Caribbean",
                    "North Africa- Middle East",
                    "Eastern Asia",
                    "Southern Asia",
                    "Western Europe- North America",
                    "Log GDP/capita",
                    "life ladder",
                     "Civil Society Repression",
                     "Generosity",
                     "South Eastern Asia",
                  "Education",
                     "Healthy life expectancy at birth"
                     ))
```

```
Error in eval(expr, envir, enclos): object 'ridge_feature_estimate_ordered' not found
```

```r
nice_table(ridge_feature_estimate_ordered_table)
```

```
Error in eval(expr, envir, enclos): object 'ridge_feature_estimate_ordered_table' not found
```

Ridge does not eliminate variables, so all variables are still listed here. It does minimize the estimate of less influential variables, though. In order of most to least important, the top 10 predictors identified by ridge regression are: civil society index, participation in democracy, social support, freedom to make life choices, government corruption index, followed by region (the Caribbean). It is interesting to note that all region variables, excluding Eastern Europe North America, social support, and freedom to make life choices are associated with a negative estimate of affect on civil society participation compared to the baseline level of Sub-Saharan Africa. This means that the civil society participation of most regions, except Eastern Europe North America, is lower than that of Sub-Saharan Africa.

Lasso Regression

```
# set seed for reproducibility
set.seed(18)

# cross validation for best l
cv_l <- cv.glmnet(x[train,], y[train], alpha = 1,
lambda = 10^seq(10, -2, length = 100), scale = TRUE)

# saving optimal lambda
bestlam_l <- cv_l$lambda.min

# calculating MSE
lasso_pred <- predict(cv_l, s = bestlam_l,
# scale x values
newx = x[test, ], scale = TRUE)
lasso_mse <- mean((lasso_pred - y[test])^2)

# fits model
lasso_mod <- glmnet(x, y, lambda = bestlam_l, scale = TRUE)

# saves coefficients
coef_l <-coef(lasso_mod)

# fits model with only important variables to get p vals
lasso_p <- lm(cspart ~ cs_index + v2x_partipdem + social_support + region +
                 lifee, cs[train,])
lasso_p <- glm(cspart ~ cs_index + v2x_partipdem + social_support + region +
                 lifee, cs[train,], family = 'gaussian')
summary(lasso_p)
```

```

Call:
glm(formula = cspart ~ cs_index + v2x_partipdem + social_support + 
    region + lifee, family = "gaussian", data = cs[train, ])

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          5.572e-01  1.441e-01   3.867 0.000184 ***
cs_index                             4.226e-01  6.723e-02   6.286 6.18e-09 ***
v2x_partipdem                        4.189e-01  1.166e-01   3.592 0.000486 ***
```

```
social_support                      -3.152e-01  1.158e-01  -2.722 0.007511 **
regionEasternEurope_PostSovietUnion -9.296e-02  2.999e-02  -3.100 0.002442 **
regionLatinAmerica                  -1.295e-01  3.679e-02  -3.519 0.000622 ***
regionNorthAfrica_MiddleEast        -4.215e-02  3.477e-02  -1.212 0.227956
regionWesternEurope_NorthAmerica     6.012e-03  4.834e-02   0.124 0.901246
regionEasternAsia                   -6.345e-02  5.167e-02  -1.228 0.222004
regionSouthEasternAsia               1.546e-03  4.235e-02   0.037 0.970939
regionSouthernAsia                  -4.070e-02  4.681e-02  -0.869 0.386491
regionThePacific                    -1.152e-01  7.284e-02  -1.581 0.116539
regionTheCarribean                  -3.596e-02  4.872e-02  -0.738 0.461936
lifee                               -8.216e-05  2.531e-03  -0.032 0.974160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.009279813)

    Null deviance: 5.2428  on 127  degrees of freedom
Residual deviance: 1.0579  on 114  degrees of freedom
AIC: -220.61

Number of Fisher Scoring iterations: 2
```

```r
# NEED TO ADD P VALUES
# prints important coefficients as a table
lasso_feature_estimate <- lasso_mod %>% tidy() %>%
  select(term, estimate)
lasso_feature_estimate_ordered <- lasso_feature_estimate %>%
  arrange(desc(abs(estimate))) #%>%
  # mutate(term = c())

## NOTE REMOVE THE INTERCEPT IN THE TABLE AND ADD P
lasso_feature_estimate_ordered_table=lasso_feature_estimate_ordered%>%
  mutate(term = c("Civil Society Index",
                  'Intercept',
                  "Democracy Participation",
                     "Social Support",
                  "Latin America",
                  "Eastern European- Post Soviet Union",
                      "The Pacific",
                   "North Africa- Middle East",
                    "Healthy life expectancy at birth"
```

```
                     ))

#print(lasso_feature_estimate_ordered)

nice_table(lasso_feature_estimate_ordered_table)
```

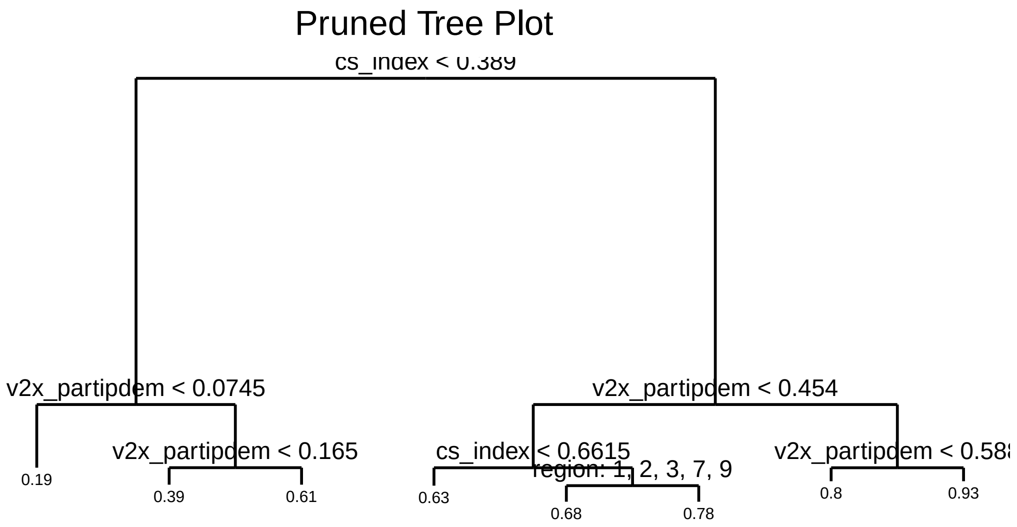
```
Error in opts_current_table(): if a label (lasso-regression) is defined, chunk option `tbl-ca
```

Lasso identified participation in democracy, social support, region, and life expectancy to be the most influential variables. The model found that Latin America, Eastern Europe and Post Soviet Union, the Pacific, and North Africa and Middle East were significantly different than the baseline Sub-Saharan Africa. The other regions, Sub Saharan Africa, Western Europe and North America, Eastern Asia, Southeastern Asia, Southern Asia, and the Caribbean, were considered to be similar enough to be categorized as one group together in terms of base level of civil society participation. This model also associates significant variables - life expectancy, and social support - with a negative effect on civil society participation.

Tree Methods

Single Decision Tree Model

```
# sets set for reproducibility
set.seed(247)

# creates tree on training data
tree_train <- tree(cspart ~ . -year -country_name, cs_num,
    subset = train)

# cross validation of tree
cv_train <- cv.tree(tree_train)

# pruning to 8 nodes per CV results
prune_train <- prune.tree(tree_train, best = 8)

# plots pruned tree
plot_tree(prune_train) +
  labs(title = "Pruned Tree Plot",
       caption = 'Regions 1, 2, 3, 7, and 9 represent Eastern Europe & Post
       Soviet Union, Latin America, North Africa and Middle East, South Eastern
       Asia, and the Pacific') +
  # Center the plot title
  theme(plot.title = element_text(hjust = 0.4))
```



Pruned Tree Plot

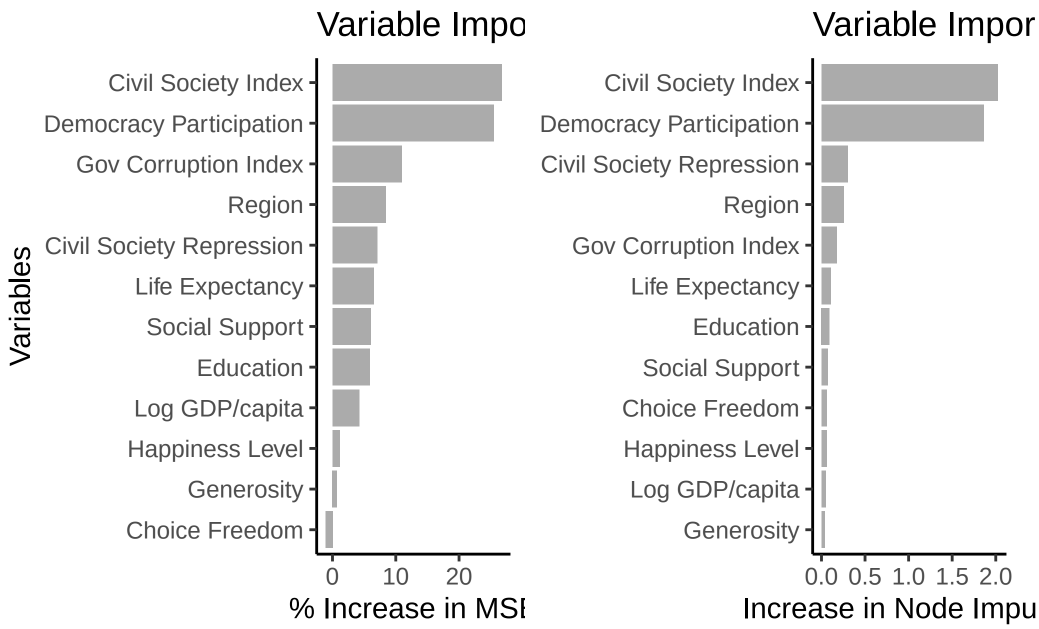
Regions 1, 2, 3, 7, and 9 represent Eastern Europe & Post
Soviet Union, Latin America, North Africa and Middle East, South Eastern
Asia, and the Pacific

```
# gets predictions and test values for mse
tree_pred_tuned <- predict(prune_train, cs_num[test,],
    type = "vector")
y_test <- y[test]

# gets mse for pruned tree
mse <- mean((tree_pred_tuned - y_test)^2)
```

The single decision tree plots the relationship between the most important predictors and the outcome. We can see that civil society index is the most influential as it is the topmost node of the tree. From there, participation in democracy and regions 1, 2, 3, 7, and 9, which represent Eastern Europe & Post Soviet Union, Latin America, North Africa and Middle East, South Eastern Asia, and the Pacific, were also identified as the most influential predictors in this model when predicting civil society participation.

Random Forest Model

```r
# sets seed for reproducibility
set.seed(286)

# sets cv parameters
train_control <- trainControl(method="cv", number = 5)

# gets grid for mtry
tune_grid <- expand.grid(mtry = 3:12)

# does training
best_forest <- train(cspart ~ . -year -country_name, data = cs[train,],
                     trControl = train_control,
                     method="rf",
                     tuneGrid = tune_grid,
                     verbose = FALSE)

# gets test for y
y_test <- y[test]

# predictions for test set with optimal mtry of 9
rf_cs <- randomForest(cspart ~ . -year -country_name, data = cs[train,],
                       mtry = 9, importance = TRUE)

yhat_rf <- predict(rf_cs, newdata = cs[test,])

# calculates MSE
rf_mse <- mean((yhat_rf - y_test)^2)

# importances
importances_cs <- importance(rf_cs) %>%
  as_tibble(rownames = "Variable")

# plot 1
p1_rf <- importances_cs %>%
  arrange(`%IncMSE`) %>%
  mutate(Variable = factor(Variable, levels = Variable)) %>%
  ggplot(aes(x = `%IncMSE`, y = Variable)) +
  geom_col(alpha = 0.5) + scale_y_discrete(
          labels = rev(c("Civil Society Index",
                         "Democracy Participation",
                         "Gov Corruption Index",
                         "Region",
                         "Civil Society Repression",
                         "Life Expectancy",
                         "Social Support",
                         "Education",
                         "Log GDP/capita",
                         "Happiness Level",
                         "Generosity",
                         "Choice Freedom"))) +
  labs(title = "Variable Importances 1",
       y = "Variables",
       x = "% Increase in MSE") +
  theme_classic()

# plot 2
p2_rf <- importances_cs %>%
  arrange(IncNodePurity) %>%
  mutate(Variable = factor(Variable, levels = Variable)) %>%
  ggplot(aes(x = IncNodePurity, y = Variable)) +
  geom_col(alpha = 0.5) + scale_y_discrete(
          labels = rev(c("Civil Society Index",
                         "Democracy Participation",
                         "Civil Society Repression",
                         "Region",
                         "Gov Corruption Index",
                         "Life Expectancy",
                         "Education",
                         "Social Support",
                         "Choice Freedom",
                         "Happiness Level",
                         "Log GDP/capita",
                         "Generosity"))) +
  labs(title = "Variable Importances 2",
       x = "Increase in Node Impurity") +
  theme_classic() +
  theme(axis.title.y = element_blank())

# side by side
p1_rf + p2_rf
```

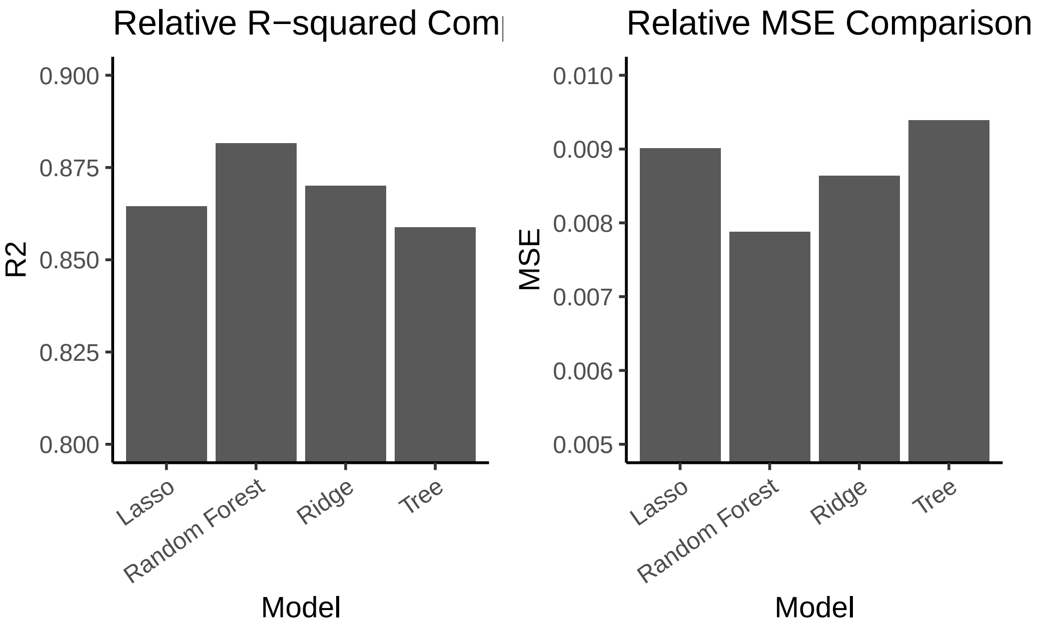


```
# prints table of most important variables
# feature names minus year, country, cspart
rf_feature_names <- colnames(cs[,c(3:6, 8:15)])

# importance scores
rf_importance_scores <- rf_cs$importance[, 1]

# combine feature names and importance scores
rf_feature_importance <- tibble(Feature = rf_feature_names,
                                'Importance Scores' = rf_importance_scores)

# get desc order for ease of interpretation
f_feature_importance_ordered <- rf_feature_importance %>%
  mutate(Feature = c("Democracy Participation",
                     "Civil Society Index",
                     "Civil Society Repression",
                     "Gov Corruption Index",
                     "Region",
                     "Life Expectancy",
                     "Education",
                     "Social Support",
                     "Log GDP/capita",
                     "Happiness Level",
                     "Choice Freedom",
                     "Generosity")) %>%
  arrange(desc(abs(rf_importance_scores)))
# Print the result
## NEED TO MAKE THE DECIMAL LESS BECAUSE WE HAVE TOO MANY .00s
nice_table(f_feature_importance_ordered)
```

```
Error in opts_current_table(): if a label (random-forests) is defined, chunk option `tbl-cap
```

The random forests model identifies civil society index and participation in democracy as the most important variables relative to other variables in the model when predicting civil society participation.

Comparing Models



Model	MSE	R-Squared
Tree	0.01	0.86
Random Forest	0.01	0.88
Lasso	0.01	0.86
Ridge	0.01	0.87

The model that performs on the test set in terms of minimizing error the best is random forests, as seen by the lowest MSE. Random forests also has the highest R-squared meaning it accounts for the most change in the predictor relative to the other models, as pictured above. This led us to conclude that random forests was the best performing model when predicting civil society participation from the data provided.

Discussion

Summary of What We Learned

Our research question regarded predictors of civil society participation. We hypothesized characteristics regarding social cohesion–such as region, presence of war, social support, and civil society index–will have similar civil society participation rates. The results of our data analysis partially supported our hypothesis, with the most important predictors identified being civil society index, region, and participation in democracy, which were seen across all models, and social support, average life expectancy, and civil society repression which was found by a minority of the models run. However, social support was shown to have a negative effect on civil society participation in 2019, which we did not expect.

Civil society index and participation in democracy were the top two predictors in all models, with region, social support, and civil society repression being identified as top predictors by some of the models. Civil society index had a positive impact on civil society participation, meaning that as elements of civil society increased so as to increase the civil society index, such as increasing the number of civil society organizations, the civil society participation would also increase. This makes sense because a higher civil society index means that there are likely more and better opportunities for people to engage in civil society, therefore increasing participation. Additionally, we must note that when considering the make-up of the civil society participation variable and the civil society index variable both consider the same piece of information in CSO Participatory Environment, which is the best description of people in civil society organizations. Both of these variables also consider three other factors, but do have an element that is in common, meaning that the relationship between the two will be in part because they share some of the same information. That said, we are unable to determine at the present how much is based on that shared information out of the entire information contained in civil society index. In future research, we would want to examine how the different aspects of civil society index (explained in the codebook) relate to civil society participation. If we were to redo this, we would consider not including this predictor because of the overlap in data, which we didn't consider to be significant given that it was one of 4 pieces of information considered. Similarly, as participation in democracy increases, so does participation in civil society. Many civil society organizations have roles in politics, so as there is more political behavior more people likely will also engage in civil society. Further study into the relationship between civil society and participation in democracy is warranted to learn more about how these two processes interact with and influence each other.

In terms of region, countries located in the Pacific, North Africa and the Middle East, Latin America, Southeastern Asia (identified only by tree model), and Eastern Europe and Post Soviet Union states were all associated with a decrease in civil society participation when compared to the other regions in the data, which are Sub Saharan Africa, Western Europe North America, Eastern Asia, Southern Asia, and the Caribbean, in both the tree model and regression models. This result was surprising because there were no regions associated with an increase in civil society participation, leading us to think that specific factors within a country may be more important than region overall. We were unable to find a commonality between all of these regions from our knowledge of these parts of the world or exploratory data analysis. Our choice of regression and tree methods does not work as well at explaining these differences as a clustering method may have, which could have illuminated the similarities between regions with similar civil society participation rates as well as other important factors. In terms of understanding why these regions are significant in more certain terms, further analysis, such as k-means clustering, is required.

The other key predictors identified were civil society repression, life expectancy, and social support, which were associated with a negative effect on civil society participation. As measures to increase repression of civil societies increase, such as the decrease of influence of civil society organizations, a decrease in civil society participation is predicted. As opportunities to participate in civil society decrease or become more difficult to access, it would make sense that people would participate less in civil society as a whole. Surprisingly, an increase in social support is associated with a decrease in civil society participation. Social support is defined as whether someone feels they have someone to turn to in a difficult situation. We hypothesized that if countries were to have more social support, it meant they would have stronger civil society networks. However, we infer that having social support may indicate that people have strong social support networks already, decreasing the need to participate in civil society organizations. More research into this relationships is needed to fully understand the effects of social support on civil society participation. Additionally, we would like to note that there are likely more relationships (interactions) between the variables than we were able to investigate, meaning social support may not truly have an overall negative effect on civil society participation as seen in the regression methods. We were not able to include interaction terms in our models, which is something that would limit our findings. Similarly, the relationship between this predictor, and other predictors, as well, and the outcome, may not have been linear, meaning that our two linear models may not have been the best representation of the data in real life. The skills of that analysis go above our abilities at this time but are worth noting as a possibility for future study.

Reliability and Validity

Reliability is the consistency or reproducibility of the results. We did our best to make results reproducible by including set seed anytime we needed random number generation. This ensures that if someone were to replicate our project they would get the same results we did. From the provided codebooks, we can tell the methods of the data collection are consistent and can be trusted. We also made sure to document our code and thinking with comments to allow for readers to follow along with our logic and conclusions. For validity, the codebooks provided thorough descriptions of how the measures were calculated, what was included, and why. These variables should have high validity. Both the V-Dem group and the data from the world happiness index are wellknown, reputable sources of data, leaving us confident that their measures are as accurate as they claim to be. As for the variables we chose, there was less validity in terms of the selected variables being good measures of civil society participation. We had minimal amounts of research when determining which variables to include in our project compared to the typical researcher in this field. This means that even though the variables measured what they claimed to, they may have had less validity in the sense that the way they were used to potentially measure civil society participation was not entirely in line with the definition of civil society or the initial meaning of the variable. Also, because of the way we added the median value for NAs of the year 2019 the analysis had less validity. This is because there are ways to better represent the true data for that country than 2019, such as replacing it with the median or mean of the missing data for the country over a certain number of years, which was above our knowledge level.

Appropriateness of Analysis

The analyses we conducted were fairly appropriate. As noted in the justification of methods in the methodology section, the regression models helped quantify the effects of predictors when others were held constant, giving us insight into which predictors had the most influence on the outcome. Lasso, especially, was helpful in this because it zeros out predictors that do not add enough to the model. Regression models are also easy to conceptualize and interpret, allowing both us and the reader to have a clearer picture of the relationship we are presenting. However, as discussed previously, the relationships in the data between the predictors and outcome may not be linear. This means that the results of the regression models may be somewhat inaccurate in their portrayal of the real relationship. However, tree models would be applicable for non-linear relationships, which is part of why we also chose a tree method to do in addition to regression. With the tree models, the tuning of hyperparameters minimized flaws of the tree model. One such issue we tried to minimize was overfitting. Because trees are locally greedy (fitting the best option for that node, not the whole tree) they run the risk of overfitting based on the data. This may result in a less accurate tree overall. Similarly, because one of the biggest risks of tree models is overfitting, if the test data is especially different than the training data it will likely do poorly. This is especially relevant as we had a smaller sample, making differences in the test vs training set more impactful. Finally, tree models are more costly in terms of computation/time even though they are very helpful in identifying predictors. However, in our case to our understanding due to the smaller number of observations we have, a tree model with random forests is not as costly as it could have been otherwise. We stand by our choice of the tree models: The single decision tree gave a strong visual depiction of the most important predictors and the relationships between them and the outcome, and random forests which identifies the most important predictors and is in general more accurate in prediction ability than the single decision tree and other methods.

Training Data Limitations

The training data we used limited our conclusions and predictions. Our data consisted of 187 observations, 70% of which were in the training set. While this is a majority of the countries of the world, it is still a limited sample. This greatly limited the generalizability of our results due to the smaller number of observations and because it removed any patterns that may have been present over time. 2019 was chosen due to a desire to understand the key factors in civil society participation as close to the present day, without the effects of Covid, as possible. However, doing a comparison over time would give us more generalizability and predicting ability by allowing us to see if predictors continued to be important over time or not. Additionally, 2019 did not have any recorded instances of civil wars or coups, meaning we were not able to assess the influence of these predictors on civil society participation. Going over time or picking a year that included these variables would have allowed us to better predict civil society participation by understanding the possible effects of these two predictors. In future works we would work to expand beyond 2019 or compare 2019 to a present post-Covid year. Working with multiple years would have allowed us to make broader generalizations about civil society participation due to the larger amount of data we could use for training. Additionally, it would have allowed us to see patterns in civil society over time and country that we were not able to see with just the 2019 data. These data were also limited in the sense that we were not able to make generalizations or predictions about regions/countries in general because we could only predict for 2019.

Looking Back

If we were to start over with this project, we would take more time to develop our questions and analysis of the data, in addition to the training set limitations discussed in the prior paragraph. Going into the project, we did not have a strong understanding of the data we were working with. This made formulating our question difficult. Doing some data exploration to learn about the relationships between explanatory variables would have helped in formulating a question. From there, being more intentional about the data we used from the beginning would have been different, as well. We had a very broad hypothesis, so narrowing down the variables we were interested in or picking a region of interest to base our analysis on would have made our data exploration and analysis more directed. In our data analysis phase, we ended up making a lot of decisions about data cleaning and data wrangling, even some about the basics of the project itself, that should have been made earlier. Choosing to focus on one year instead of multiple was a decision we do not regret, but coming to this choice at the beginning of the project would have greatly changed our data exploration and cleaning phases, as we had to approach them from a different perspective and basically redo them to do data analysis. Furthermore, learning more about the predictors themselves is a change we would make. We did not assess for covariance or any other sort of relationship between predictors. We also did not realize that some predictors we chose for our model would simply be NA for all observations, leading us to remove it later on. Additionally, the late inclusion or dropping of predictors, which proved to be tedious, would have been avoided had we taken more time to fully explore the data in the beginning. In summary, if we were to do this project again, we would do a more in-depth exploration of the data itself, think more carefully about the feasibility of our research question, then be more intentional about the data exploration and visualization we do.