The Inequality Project

Oliver Hagaseth Mydske - University of Oslo

19.05.2022

Abstract

Does economic equality foster democracy, and does inequality pave the way for autocracy? This paper examines 194 countries, to see if a larger gap between the rich and the poor, affects the quality of democracy. Multiple measurements of both equality and democracy were analysed using economic regression models, comparison of means and contingency tables. It was found that equal countries were significantly more democratic than unequal ones. Equality was strongly correlated with democracy, and the effect persisted over time. However, this pattern did not hold when economic equality was measured after tax. It is the underlying allocation of income in the economy, before redistribution is taken into account, which is predictive of the state of society's democracy. This imply that tax systems in both democracies and autocracies alike, alleviate income inequality to a certain degree. The economic regression models fell short of fully explaining the contemporary wave of autocratization in the world, suggesting that important non-economic factors are behind recent trends.

Contents

Acknowledgments

I would like to express my gratitude to my lecturers, Philipp Christian Broniecki, Felix Haass and Håvard Strand, for untangling the mystical nature of R. I would also like to thank my friends Kristine Helskog, Annika Klavenes Syverstad and Fredrik Rinde Thorstensen, who gave me inspiration and feedback throughout the project.

1 Introduction

Hypotheses

- H1: Democracy can be explained by economic factors, chief among which is economic equality.
- H2: Equal countries are democracies, and unequal countries are autocracies.
- H3: Countries that are increasingly economically equal, experience democratization, while countries that are decreasingly equal, experience autocratization.
- H4: More countries are becoming increasingly equal and authoritarian.

Packages

```
# Data packages

library(vdemdata)

# Statistical packages

library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:ggcorrplot':
##
##     cor_pmat
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```r
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
# Diagram and map packages

library(highcharter)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(naniar)
library(rnaturalearth)
library(sf)
```

```
## Linking to GEOS 3.9.1, GDAL 3.2.1, PROJ 7.2.1; sf_use_s2() is TRUE
```

```r
# Tidy packages

library(countrycode)
library(ggthemes)
library(scales)
library(ggrepel)
library(knitr)
library(kableExtra)
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks rstatix::filter(), stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

Word count

The word count is 4752.

```
academicWriteR::count_words("Inequality-project.Rmd")
```

```
## [1] 4885
```

2 Importing data

In the very first part of my analysis, I download the necessary data to sufficiently test my hypotheses.

Democracy data

To measure the concept of democracy, I chose the Democracy Index developed by the Economist Intelligence Unit, and the Liberal Democracy Index developed by the Varieties of Democracy Institute. The former was selected because it represents the effort of a single expert group to qualitative measure democracy, and for its widespread use in public discussion. The latter was selected because it represents a thorough quantitative effort from multiple scholars and teams worldwide, to measure democracy based on Robert Dahl's understanding of democracy. A such, I hope the two different measurements will nicely capture the differences in which the quantitative/qualitative approaches influence the final analysis.

```
raw_data_vdem <- vdem   #Aquiering VDEM Data


raw_data_eiu <- read_excel(   #Aquiering EIU Data
  path = "../Data/_EIU-Democracy Indices - Dataset - v4.xlsx",
  sheet = 4)

# https://www.gapminder.org/data/documentation/democracy-index/
# Link to EIU dataset
```

Economic equality data

To measure the concept of economic equality, I chose the Gini Index developed by the Italian statistician Corrado Gini. The original Gini Index measures income distribution across a population by assigning a value between 0 to 100, with 0 representing perfect equality and 100 representing perfect inequality (1 person owning everything). However, in an effort to make the data I present more understandable to the reader, I have reversed the Gini Index. This means that 1 will represent perfect equality, while 0 will represent perfect inequality. This has three advantages compared to the original: (1) avoiding confusing double negatives caused by the in- prefix in inequality, (2) making it readily comparable to the democracy indicator, which assign perfect democracy a value of 1, (3) avoiding the need to reverse the axis, in order to make a scatter plot going from bottom left to top right. I have used Our World in Data to measure the Gini index after tax, and conversely, World Inequality Database, to measure Gini Index before tax. I hope the two different measurements will nicely capture the differences in which the tax system influences the final analysis.

```
raw_data_wb <- read_excel(path = "../Data/WDIEXCEL.xlsx")   #Aquiering WB Data
country_codes <- read_excel(path = "../Data/WDIEXCEL.xlsx",
                            sheet = "Country")

# https://databank.worldbank.org/source/world-development-indicators
# Link to WB dataset

raw_data_owid <- read.csv("../Data/economic-inequality-gini-index.csv")

# https://ourworldindata.org/income-inequality
# Link to OWID dataset

raw_data_wid <- read_excel("../Data/WID_Data_02032022-194319.xls")

# https://wid.world/data/#countrytimeseries/gptinc_p0p100_z/NZ/1921/2021/eu/k/p/yearly/g
# Link to WID dataset

country_regions <- country_codes %>%   #Useful for seperating countries and regions
  select(`Country Code`, Region)

country_names <- country_codes %>%   #Will come in handy when I make maps
  select(`Short Name`, Region) %>%
  drop_na() %>%
  select(-Region)
```

3 Manipulating data

In the third part of the analysis, I manipulate the downloaded data in order to make it tidy. This step mostly consists of finding the right country-year observations, filtering out regions, renaming variable names, and making sure all variables are of the right level of measurement.

Democracy data

VDEM (Varieties of Democracy Institute)

```
# Converting the democracy indicator to a 0-100 index

raw_data_vdem$v2x_libdem <- raw_data_vdem$v2x_libdem*100

# Liberal Democracy Index

data_vdem <- raw_data_vdem %>%
  filter(year %in%  c(1960:2020)) %>%   #Relevant years
  mutate(year = as.character(year)) %>%
  select(country_name, year, v2x_libdem) %>%   #Interesting indicators
  rename(country = country_name,
         libdem = v2x_libdem) %>%
```

```r
  mutate(id = countrycode(country,   #Mutating both ways to maximise results
                          origin = "country.name",
                          destination = "iso3c"),
         id = replace(id, country == "Taiwan", "TWN"),
         country = countrycode(id,
                               origin = "iso3c",
                               destination = "country.name"),
         Continent = countrycode(country,
                                 origin = "country.name",
                                 destination = "un.region.name")) %>%
  pivot_wider(names_from = year,   #Necessary for using rowMeans
              values_from = libdem)

# Data wrangling in order to aquire two 3-year periods

data_vdem <- data_vdem[-c(19,122), ] #Removing glitch

data_vdem[4:64] = lapply(data_vdem[4:64],   #Preparing to converting NULL to NA
                         FUN = function(x) {as.character(x)}
                         )

data_vdem[data_vdem == "NULL"] <- NA   #Converting NULL to NA

data_vdem[4:64] = lapply(data_vdem[4:64],   #Converting back to numeric
                         FUN = function(x) {as.numeric(x)}
                         )

data_vdem$past <- rowMeans(subset(data_vdem,   #2006/2007/2008
                                  select= c(`2006`, `2007`, `2008`),
                                  na.rm = TRUE))

data_vdem$present <- rowMeans(subset(data_vdem,   #2018/2019/2020
                                     select= c(`2018`, `2019`, `2020`),
                                     na.rm = TRUE))

data_vdem <- data_vdem %>%   #Making it tidy once more
  pivot_longer(c(`1960`:`present`),
               names_to = "time",
               values_to = "libdem")
```

EIU (The Economist Intellegence Unit)

```r
# Democracy Index

data_eiu <- raw_data_eiu %>%
  filter(time %in%  c(2006:2020)) %>%   #Relevant years
  mutate(time = as.character(time)) %>%
  select(name, time, `Democracy index (EIU)`) %>%   #Interesting indicators
  rename(country = name,
```

```
         year = time,
         demindex = `Democracy index (EIU)`) %>%
  mutate(id = countrycode(country,    #Mutating both ways to maximise results
                          origin = "country.name",
                          destination = "iso3c"),
         id = replace(id, country == "Taiwan", "TWN"),
         country = countrycode(id,
                               origin = "iso3c",
                               destination = "country.name"),
         Continent = countrycode(country,
                                 origin = "country.name",
                                 destination = "un.region.name")) %>%
  pivot_wider(names_from = year,    #Necessary for using rowMeans
              values_from = demindex)
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```
data_eiu$past <- rowMeans(subset(data_eiu,    #2006/2007/2008
                                 select= c(`2006`, `2007`, `2008`),
                                 na.rm = TRUE))

data_eiu$present <- rowMeans(subset(data_eiu,    #2018/2019/2020
                                    select= c(`2018`, `2019`, `2020`),
                                    na.rm = TRUE))

data_eiu <- data_eiu %>%    #Making it tidy once more
  pivot_longer(c(`2006`:`present`),
               names_to = "time",
               values_to = "demindex")
```

Economic equality data

WB (World Bank)

```
# GDP per capita, total population and dependence on natural resources

data_wb <- raw_data_wb %>%
  left_join(country_regions, by = "Country Code") %>%
  filter(!is.na(Region),    #Removing non-country observations
         `Indicator Code` %in% c("NY.GDP.PCAP.PP.KD",    #GDP per capita
                                 "SP.POP.TOTL",    #Population
                                 "NY.GDP.TOTL.RT.ZS")) %>%  #Natural resources
  select(-`Indicator Name`) %>%    #Avoiding confusion with indicator code
  pivot_longer(`1960`:`2020`,    #Relevant years
               names_to = "year",
               values_to = "values") %>%
  pivot_wider(names_from = `Indicator Code`,
              values_from = values) %>%
```

```r
  rename(country = `Country Name`,
         id = `Country Code`,
         gdpcapita = `NY.GDP.PCAP.PP.KD`,
         population  = `SP.POP.TOTL`,
         resources = `NY.GDP.TOTL.RT.ZS`) %>%
  mutate(id = countrycode(country,    #Mutating both ways to maximise results
                          origin = "country.name",
                          destination = "iso3c"),
         id = replace(id, country == "Micronesia (country)", "FSM"),
         country = countrycode(id,
                               origin = "iso3c",
                               destination = "country.name"),
         Continent = countrycode(country,
                                 origin = "country.name",
                                 destination = "un.region.name")
         ) %>%
  select(country, id, Continent, year, gdpcapita, population, resources)
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```r
# Data wrangling in order to aquire two 3-year periods

data_wb <- data_wb %>%
  pivot_wider(names_from = year,
              values_from = c(gdpcapita, population, resources))
```

```
## Warning: Values from `gdpcapita`, `population` and `resources` are not uniquely identified; out
## * Use `values_fn = list` to suppress this warning.
## * Use `values_fn = {summary_fun}` to summarise duplicates.
## * Use the following dplyr code to identify duplicates.
##   {data} %>%
##     dplyr::group_by(country, id, Continent, year) %>%
##     dplyr::summarise(n = dplyr::n(), .groups = "drop") %>%
##     dplyr::filter(n > 1L)
```

```r
data_wb <- data_wb[-40, ] #Removing gltich

data_wb[4:186] = lapply(data_wb[4:186], #rowMeans is only applicable to numeric
                        FUN = function(x) {as.numeric(x)}
                        )

data_wb$gdpcapita_past <-
  rowMeans(subset(data_wb,
                  select= c(gdpcapita_2006, gdpcapita_2007, gdpcapita_2008),
                  na.rm = TRUE))

data_wb$gdpcapita_present <-
  rowMeans(subset(data_wb,
                  select = c(gdpcapita_2018, gdpcapita_2019, gdpcapita_2020),
```

```r
                  na.rm = TRUE))

data_wb$population_past <-
  rowMeans(subset(data_wb,
                  select = c(population_2006, population_2007, population_2008),
                  na.rm = TRUE))

data_wb$population_present <-
  rowMeans(subset(data_wb,
                  select = c(population_2018, population_2019, population_2020),
                  na.rm = TRUE))

data_wb$resources_past <-
  rowMeans(subset(data_wb,
                  select = c(resources_2006, resources_2007, resources_2008),
                  na.rm = TRUE))

data_wb$resources_present <-
  rowMeans(subset(data_wb,
                  select = c(resources_2006, resources_2007, resources_2008),
                  na.rm = TRUE))

data_wb <- data_wb %>%   #Making it tidy once more
  pivot_longer(c(`gdpcapita_1960`:`gdpcapita_2020`,
                 gdpcapita_past, gdpcapita_present),
               names_to = "time",
               names_prefix = "gdpcapita_",
               values_to = "gdpcapita") %>%
  pivot_longer(c(`population_1960`:`population_2020`,
                 population_past, population_present),
               names_to = "time2",
               names_prefix = "population_",
               values_to = "population") %>%
  pivot_longer(c(`resources_1960`:`resources_2020`,
                 resources_past, resources_present),
               names_to = "time3",
               names_prefix = "resources_",
               values_to = "resources")

# Combining the three time variables and selecting only one time variable

data_wb <- data_wb[data_wb$time==data_wb$time2 & data_wb$time2==data_wb$time3,]

data_wb <- data_wb %>%
  select(country, id, Continent, time, gdpcapita, population, resources)
```

OWID (Our World in Data)

```r
# Reverse Gini index to a 0-100 scale

raw_data_owid$Gini.index <- 100-raw_data_owid$Gini.index

# Gini Index after tax

data_owid <- raw_data_owid %>%
  filter(Year %in%  c(1981:2021),   #Relevant years
         !(str_detect(Entity, "Rural"))) %>% #Removing non-complete observations
  mutate(Year = as.character(Year)) %>%
  rename(gini = Gini.index,
         country = Entity,
         year = Year) %>%
  mutate(id = countrycode(country,   #Mutating both ways to maximise results
                          origin = "country.name",
                          destination = "iso3c"),
         id = replace(id, country == "Taiwan", "TWN"),
         country = countrycode(id,
                               origin = "iso3c",
                               destination = "country.name"),
         Continent = countrycode(country,
                                 origin = "country.name",
                                 destination = "un.region.name")) %>%
  select(-Code) %>%
  pivot_wider(names_from = year, values_from = gini)

data_owid <- data_owid[-c(139), ] #Removing gltich

data_owid[4:42] = lapply(data_owid[4:42],   #Preparing to converting NULL-values
                         FUN = function(x) {as.character(x)}
                         )

data_owid[data_owid == "NULL"] <- NA   #Converting NULL to NA

data_owid[4:42] = lapply(data_owid[4:42],   #Preparing to use rowMeans
                         FUN = function(x) {as.numeric(x)}
                         )

data_owid$past <- apply(X = subset(data_owid,
                                   select = c(`2006`, `2007`, `2008`)),
                        MARGIN = 1,   #Row-wise operations
                        FUN = mean,   #Finding the mean
                        na.rm = TRUE)   #Ignore NAs

data_owid$present <- apply(X = subset(data_owid,
                                      select = c(`2018`, `2019`)), #Missing 2020
                           MARGIN = 1,   #Row-wise operations
                           FUN = mean,   #Finding the mean
```

```
                        na.rm = TRUE)   #Ignore NAs

data_owid <- data_owid %>%
  pivot_longer(c(`1981`:`present`),
               names_to = "time",
               values_to = "gini_after_tax")
```

WID (World Inequality Database)

```
# Gini Index before tax

names(raw_data_wid)[3:124] <- paste(1900:2021) #Clearing the year labels

raw_data_wid[3:124] = lapply(raw_data_wid[3:124],   #Preparing to use rowMeans
                             FUN = function(x) {as.numeric(x)}
                             )

raw_data_wid$past <- rowMeans(subset(raw_data_wid,
                                  select= c(`2006`, `2007`, `2008`),
                                  na.rm = TRUE))

raw_data_wid$present <- rowMeans(subset(raw_data_wid,
                                        select= c(`2018`, `2019`, `2020`),
                                        na.rm = TRUE))

data_wid <- raw_data_wid %>%
  rename(country = Country) %>%
  pivot_longer(c(`1900`:`present`),
               names_to = "time",
               values_to = "gini_before_tax") %>%
  filter(time > 1959,
         !(str_detect(country, #Removing non-complete country-observations
              "(at market exchange rate)|&|Rural|Urban|Other|West|East"))) %>%
  mutate(id = countrycode(country,   #Mutating both ways to maximise results
                          origin = "country.name",
                          destination = "iso3c"),
         id = replace(id, country == "Taiwan", "TWN"),
         country = countrycode(id,
                               origin = "iso3c",
                               destination = "country.name"),
         Continent = countrycode(country,
                                 origin = "country.name",
                                 destination = "un.region.name")) %>%
    select(country, id, Continent, time, gini_before_tax)
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So

## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```r
# Reverse Gini index to a 0-100 scale

data_wid$gini_before_tax <- (1-data_wid$gini_before_tax)*100

# Removing South Africa which accidently got included twice

data_wid <- data_wid[-c(9985:10048), ]
```

Merging data

Unifying data sets

Last, but not least, I merge together the critical variables from the five different data sets, into a single data frame. This data frame will be my focus of attention for the rest of the analysis.

```r
# Merging data

merged_data1 <- data_owid %>%   #Economic equality after tax
  full_join(data_wid,   #Economic equality before tax
            by = c("country", "id", "Continent", "time")) %>%
  full_join(data_wb,   #GDP per capita, population and natural resources
            by = c("country", "id", "Continent", "time")) %>%
  full_join(data_vdem,   #Liberal Democracy Index
            by = c("country", "id", "Continent", "time")) %>%
  full_join(data_eiu,   #Democracy Index
            by = c("country", "id", "Continent", "time")) %>%
  drop_na(country) %>%   #Dropping missing countries (ie. regions)
  na_if("NaN") %>%   #Converting Not a Number (NaN) to Not Available (NA)
  filter(!if_all(c(5:9), is.na),    #Removing all NA-values
         !(str_detect(country, "Palestinian")), #Removing non-sovereign states:
         !(str_detect(country, "Aruba")),       #Aruba
         !(str_detect(country, "Sint")),        #Sint maarten
         !(str_detect(country, "SAR")),         #Hong Kong and Macao
         !(str_detect(country, "Turks")),       #Turks and Caicos Islands
         !(str_detect(country, "Bermuda")),     #Bermuda
         !(str_detect(country, "Puerto")),      #Puerto Rico
         !(str_detect(country, "Caledonia")),   #New Caledonia
         !(str_detect(country, "Isle")),  #Isle of Man
         !(str_detect(country, "Gibraltar")),   #Gibraltar
         !(str_detect(country, "Polynesia")),   #French Polynesia
         !(str_detect(country, "Greenland")),   #Greenland
         !(str_detect(country, "Virgin")),      #British and US Virgin Islands
         !(str_detect(country, "American")),    #American Samoa
         !(str_detect(country, "Guam")),        #Guam
         !(str_detect(country, "Saint")),       #Saint Maartin
         !(str_detect(country, "Faroe")),       #Faroe Islands
         !(str_detect(country, "Mariana")),     #Mariana Islands
         !(str_detect(country, "Cayman")),      #Cayman Islands
         !(str_detect(country, "Curaçao"))      #Curaçao
         ) %>%
```

```r
  arrange(desc(country))   #Sorting alphabetically

# Changing continent variable

merged_data1[1533:1596, 3] <- "Asia"       #Taiwan is part of Asia
merged_data1[958:1021, 3] <- "Europe"      #Turkey is part of Europe
merged_data1[8300:8363, 3] <- "Europe"     #Georgia is part of Europe
merged_data1[9514:9577, 3] <- "Europe"     #Cyprus is part of Europe
merged_data1[11688:11751, 3] <- "Europe"   #Azerbaijan is part of Europe
merged_data1[11880:11943, 3] <- "Europe"   #Armenia is part of Europe

# Changing country name variable

merged_data1[4851:4914, 1] <- "Myanmar"
merged_data1[5298:5369, 1] <- "Micronesia"

# Adding ordinal equality variables

merged_data1 <- merged_data1 %>%
  mutate(equality_after_tax = cut(gini_after_tax,
                                breaks = c(0, 65, Inf),
                                labels = c("Unequal", "Equal")),
         equality_before_tax = cut(gini_before_tax,
                                 breaks = c(0, 45, Inf),
                                 labels = c("Unequal", "Equal")),
         regime = case_when(gini_after_tax > 65 & libdem > 35 ~ 1,
                            gini_after_tax > 65 & libdem < 35 ~ 2,
                            gini_after_tax < 65 & libdem > 35 ~ 3,
                            gini_after_tax < 65 & libdem < 35 ~ 4)
         )

# Adding socialism as control variable

merged_data1$socialism=0 #All non-socialist countries

merged_data1$socialism <-
  ifelse(grepl(
    c("Armenia|Azerbaijan|Belarus|China|Cuba|Russia|Kazakhstan|Mongolia|Kyrgyzstan|North Korea|Uzb
    merged_data1$country), 1, 0)

# https://commons.wikimedia.org/wiki/File:Socialist_states_by_duration_of_existence.png
# Link to classification of socialist countries.
# A country is considered socialist if a socialist regime has been in power 60 or more years.

# Merged data for 2019

correlation <- merged_data1 %>%
  filter(time == "2019")

nrow(correlation)   #A total of 194 countries is covered in the analysis
```

```
## [1] 194
```

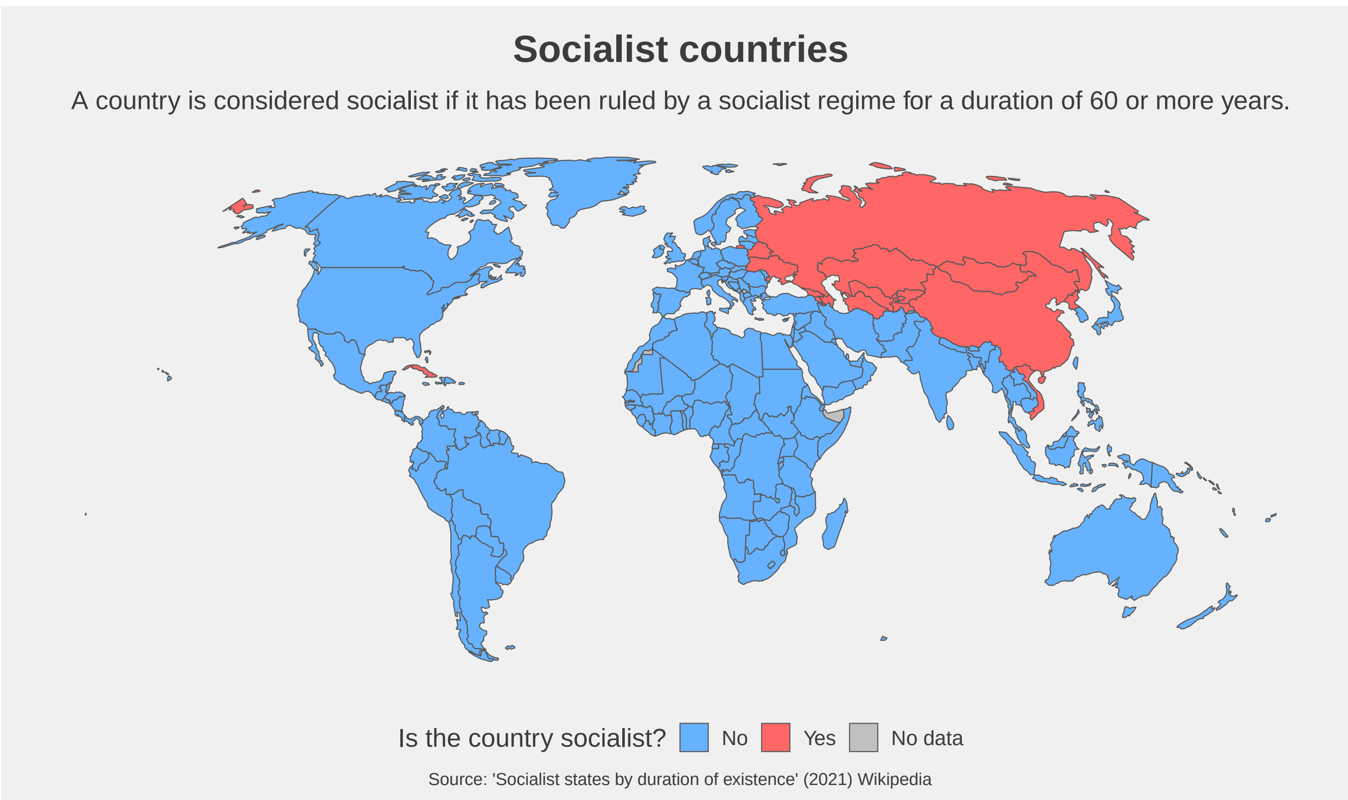
Map data

```
# Map data

raw_map_data <- ne_countries(scale = 'small',    #Removing small islands
                             type = 'sovereignty',    #Sovereignty counts
                             returnclass = 'sf')

map_data1 <- raw_map_data %>%
  select(sovereignt, geometry) %>%
  filter(sovereignt != "Antarctica") %>%    #Removing Antarctica
  rename(country = sovereignt) %>%
  mutate(id = countrycode(country,    #Making names similair to merged_data1
                          origin = "country.name",
                          destination = "iso3c")) %>%
  left_join(correlation, by = "id") %>%
  mutate(regime = case_when(    #Requirement to use numerical values in sf-maps
    regime == 1 ~ "Equal democracy",
    regime == 2 ~ "Equal autocracy",
    regime == 3 ~ "Unequal democracy",
    regime == 4 ~ "Unequal autocracy")) %>%
  mutate(socialism = case_when(    #Socialism as a dummy variable
    socialism == 0 ~ "No",
    socialism == 1 ~ "Yes"))
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```
# Map colors

map1_colors <- c("No" = "#66B2FF",
                 "Yes" = "#FF6666",
                 "No data" = "#C0C0C0")

# Map making

map1 <- ggplot(data = map_data1) +
  geom_sf(aes(fill = socialism), lwd = 0.3) +    #Reducing country border width
  scale_fill_manual(values = map1_colors,    #Costum colors
                    na.value = "#C0C0C0",
                    labs(fill = "Is the country socialist?"))+
  coord_sf(crs ="+proj=robin +lon_0=0 +x_0=0    #A better map projection
           +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")+
  labs(title = "Socialist countries",
       subtitle = "A country is considered socialist if it has been ruled by a socialist regime fo
       caption = "Source: 'Socialist states by duration of existence' (2021) Wikipedia")+
  theme_fivethirtyeight()+    #Pleasent theme
  theme(panel.grid.major = element_blank(),    #Removing longditude and latitude
```

```
        axis.text = element_blank(),    #No unnecessary titles
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(2,0,2,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 12))

# Showing and saving map

map1
```




```
ggsave("map1.png", width = 10, height = 6)
```

4 Exploring data

In part four, I explore important characteristics of the finalised data set. To get a sense of what kind of data I am working with, I will identify the central tendencies, the missing values, and the dispersion in the different variables. I provide visualizations of the results.

Missing values

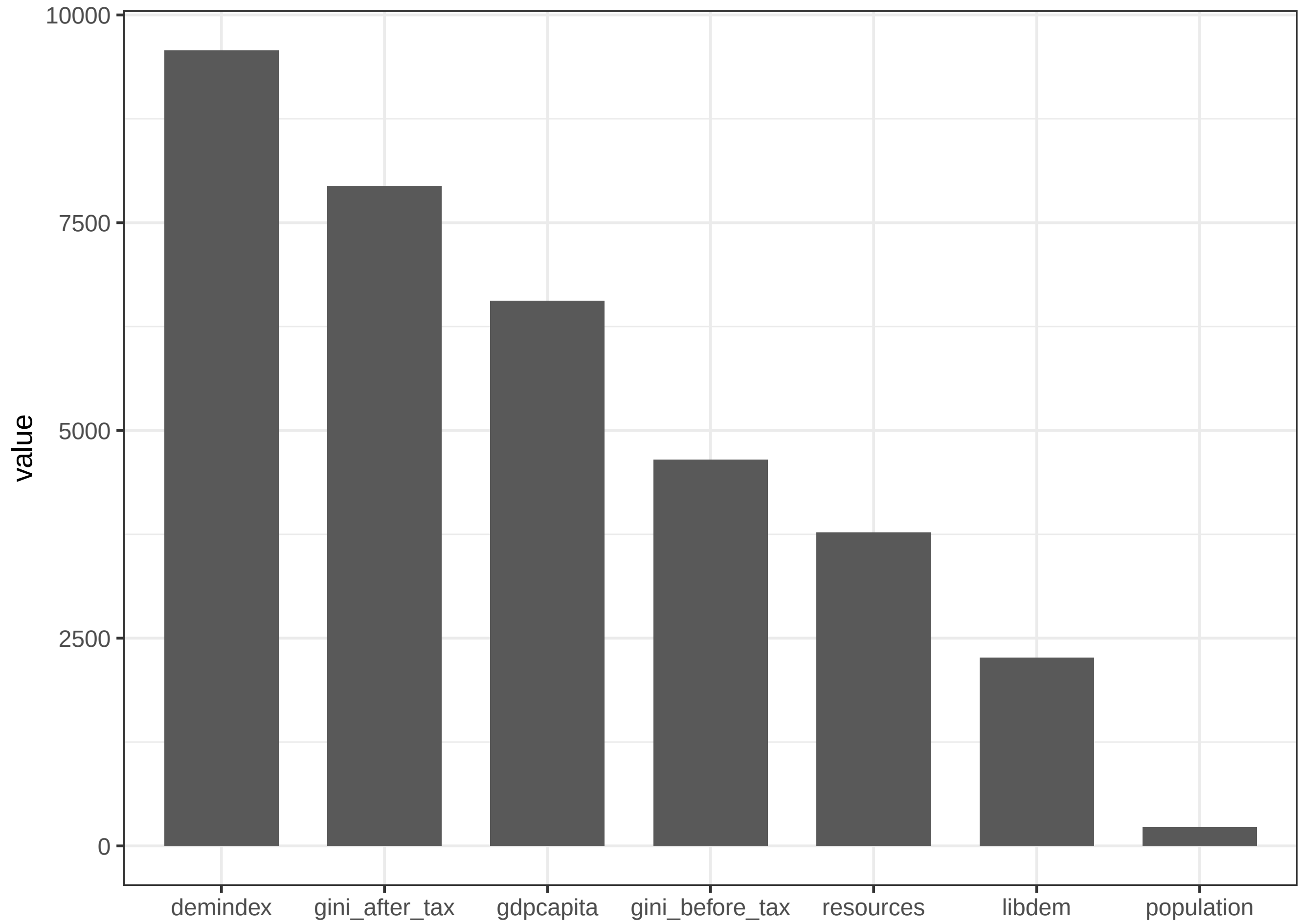
Missing values and outliers are particularly useful to identify. For instance, it is difficult to conduct a proper analysis with too many missing variables. The risks of committing both type I (false positive) and type II errors (false negative), are greatly increased when the sample size is lower due to missing values. To avoid concluding that results are statistically significant when, in reality, they came about by chance or unrelated factors, I will use a significance level of 5% in the rest of my analysis.

```
# Summary

summary(correlation)
```

```
##    country                id              Continent             time
##  Length:194         Length:194         Length:194         Length:194
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  gini_after_tax  gini_before_tax   gdpcapita          population
##  Min.   :36.97   Min.   :25.35   Min.   :   751.7   Min.   :1.076e+04
##  1st Qu.:57.62   1st Qu.:39.18   1st Qu.:  4718.7   1st Qu.:2.038e+06
##  Median :63.24   Median :43.12   Median : 13527.4   Median :8.967e+06
##  Mean   :61.97   Mean   :43.86   Mean   : 20533.8   Mean   :3.977e+07
##  3rd Qu.:67.26   3rd Qu.:49.43   3rd Qu.: 29517.1   3rd Qu.:2.875e+07
##  Max.   :75.37   Max.   :62.28   Max.   :116518.3   Max.   :1.408e+09
##  NA's   :33      NA's   :22      NA's   :11         NA's   :2
##    resources           libdem          demindex     equality_after_tax
##  Min.   : 0.0000   Min.   : 0.90   Min.   :10.80   Unequal:98
##  1st Qu.: 0.2582   1st Qu.:16.80   1st Qu.:33.00   Equal  :63
##  Median : 1.2889   Median :38.00   Median :56.50   NA's   :33
##  Mean   : 4.9987   Mean   :41.06   Mean   :54.46
##  3rd Qu.: 5.0595   3rd Qu.:63.20   3rd Qu.:71.70
##  Max.   :47.8687   Max.   :88.30   Max.   :98.70
##  NA's   :15        NA's   :21      NA's   :29
##  equality_before_tax     regime         socialism
##  Unequal:98          Min.   :1.000   Min.   :0.00000
##  Equal  :74          1st Qu.:1.250   1st Qu.:0.00000
##  NA's   :22          Median :3.000   Median :0.00000
##                      Mean   :2.617   Mean   :0.08247
##                      3rd Qu.:4.000   3rd Qu.:0.00000
##                      Max.   :4.000   Max.   :1.00000
##                      NA's   :40
```

```
# Bar plot of missing values in each variable

ggplot(data = merged_data1 %>%
         summarise_all(~ sum(is.na(.))) %>%   #Adding together NAs
         pivot_longer(cols = c(5:11),
```

```
                       names_to = "variable",
                       values_to = "value"),
       aes(x = reorder(variable, -value),    #Sorting by number og NAs
           y = value))+
  geom_bar(width = 0.7,
           stat = "identity")+
  theme_bw()+    #Pleasent theme
  theme(axis.title.x = element_blank())    #Removing x-axis
```




```
# Countries with no data on economic equality

correlation %>%    # 42 countries missing data from economic equality
  filter(is.na(gini_after_tax)|is.na(gini_before_tax)) %>%
  select(country, gini_after_tax, gini_before_tax) %>%
  arrange(gini_after_tax)
```

```
## # A tibble: 42 x 3
##    country         gini_after_tax gini_before_tax
##    <chr>                    <dbl>           <dbl>
##  1 St. Lucia                 48.8              NA
##  2 Tuvalu                    60.9              NA
##  3 Samoa                     61.3              NA
```

```
##   4 Vanuatu                    62.4           NA
##   5 Tonga                      62.4           NA
##   6 Solomon Islands            62.9           NA
##   7 Kiribati                   63.0           NA
##   8 Fiji                       63.3           NA
##   9 Nauru                      65.2           NA
## 10 Venezuela                  NA             39.6
## # ... with 32 more rows
```

```
# Number of missing values in each row

correlation$missing <- rowSums(is.na(correlation[,5:10]))

missing_data <- correlation %>%   #Number of NAs by country-observation
  select(country, missing) %>%
  filter(missing > 0) %>%
  arrange(-missing)

missing_data
```

```
## # A tibble: 50 x 2
##    country                  missing
##    <chr>                      <dbl>
##   1 Monaco                         5
##   2 Liechtenstein                  5
##   3 San Marino                     4
##   4 Micronesia                     4
##   5 Marshall Islands               4
##   6 Eritrea                        4
##   7 Andorra                        4
##   8 Venezuela                      3
##   9 Taiwan                         3
## 10 St. Vincent & Grenadines       3
## # ... with 40 more rows
```

```
# Word cloud of missing countries

text <- lapply(missing_data, rep, missing_data$missing)   #Converting to text

text_data <- data.frame(text = text,
                        freq = text$missing,
                        stringsAsFactors = FALSE)

text_aggregated <- aggregate(freq ~ ., data = text_data, sum)

# png("wordcloud.png", width = 7, height = 7, units='in', res = 500)
# Higher resolution word cloud by removing the # in front of png("...")

word_cloud <- wordcloud::wordcloud(words = text_aggregated$text.country,
          freq = text_aggregated$freq,
```

```
          min.freq = 1,
          max.words = 200,
          random.order = FALSE,
          rot.per = 0.35,
          scale = c(3.1, 0.2),
          colors = RColorBrewer::brewer.pal(8, "Dark2"))
```

There is a worrisome amount of missing country-observations in some of the data sets. Gini after tax has a lot more missing countries, compared to Gini before tax, 33 in total. That is one of the main reasons I will focuse my attention on economic equality before taxation in this analysis. Luckily, it seems that most of the countries missing the most amount of data, are small island nations in the pacific, or other micro countries. Monaco and Lienchtenstein stand out, for being almost completely absent in the data sets. Their absence should not have an substantial impact on the analysis.

Density plots

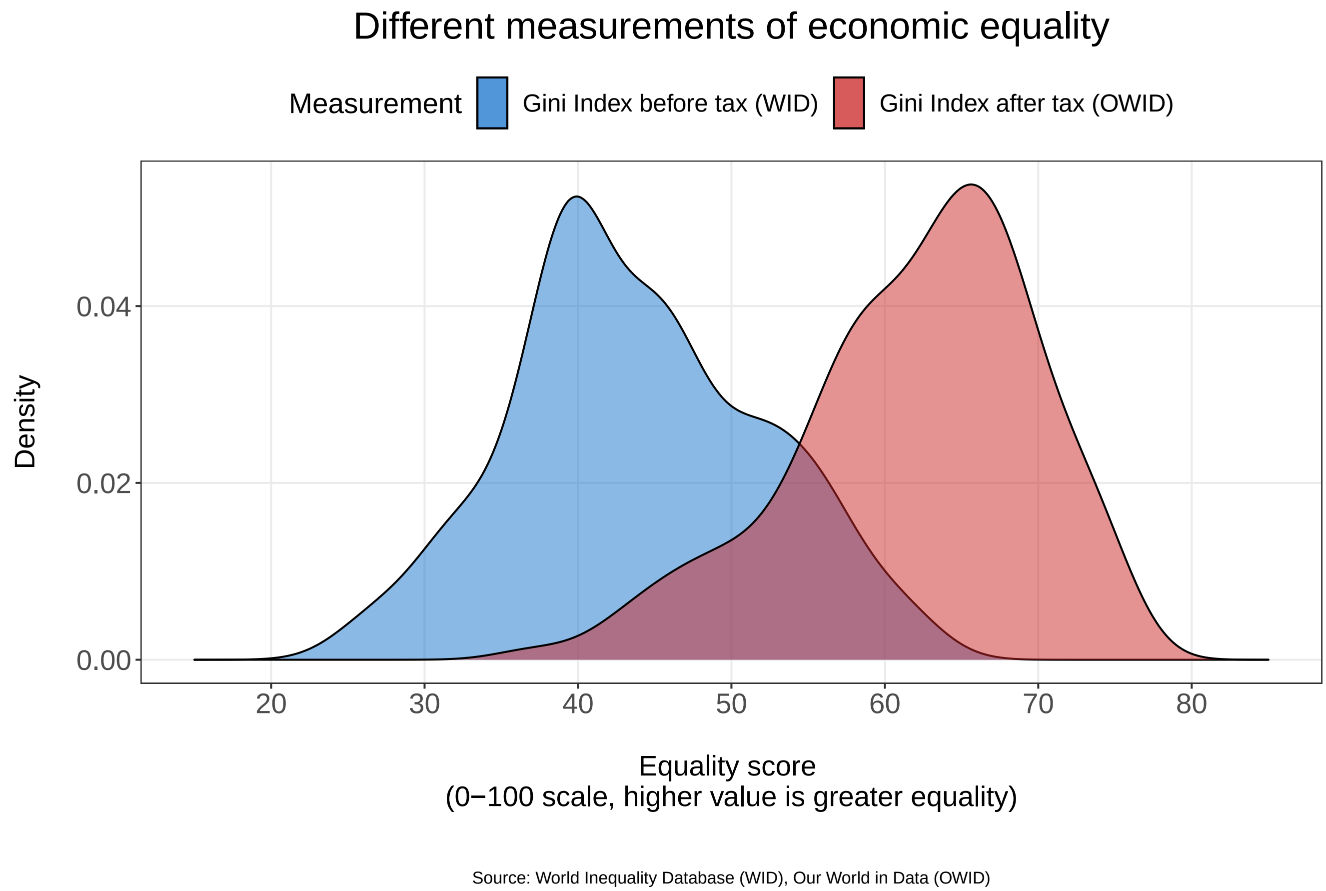
It's time to wrap my head around the differences between my main variables, and how they are going to shape the analysis. I use density plots to compare and contrast the two different measures of democracy and economic equality with each other.

```
# Colors
```

```r
equality_colors <- c("Gini Index before tax (WID)" = "dodgerblue3",
                     "Gini Index after tax (OWID)" = "firebrick3")

democracy_colors <- c("Liberal Democracy Index (VDEM)" = "dodgerblue3",
                      "Democracy Index (EIU)" = "firebrick3")

# Equality density plot


density_equality <- ggplot(data = correlation)+
  geom_density(aes(x = gini_before_tax,
                   fill = "Gini Index before tax (WID)"),
               alpha = 0.5)+
  geom_density(aes(x = gini_after_tax,
                   fill = "Gini Index after tax (OWID)"),
               alpha = 0.5)+
  scale_fill_manual(values = equality_colors,
                     labs(fill = "Measurement"))+
  scale_x_continuous(limits = c(15, 85),
                     breaks = seq(0, 100, by = 10))+
  labs(title = "Different measurements of economic equality",
       caption = "\nSource: World Inequality Database (WID), Our World in Data (OWID)\n",
       x = "\nEquality score \n(0-100 scale, higher value is greater equality)\n",
       y = "\nDensity\n") +
  theme_bw()+
  theme(legend.position = "top",   #Legend on top
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 13),
        legend.key.height = unit(1, 'cm'),
        panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5,
                                  size = 20,
                                  margin=margin(20,0,10,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 12,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5)
        )

density_equality
```

```
## Warning: Removed 22 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 33 rows containing non-finite values (stat_density).
```



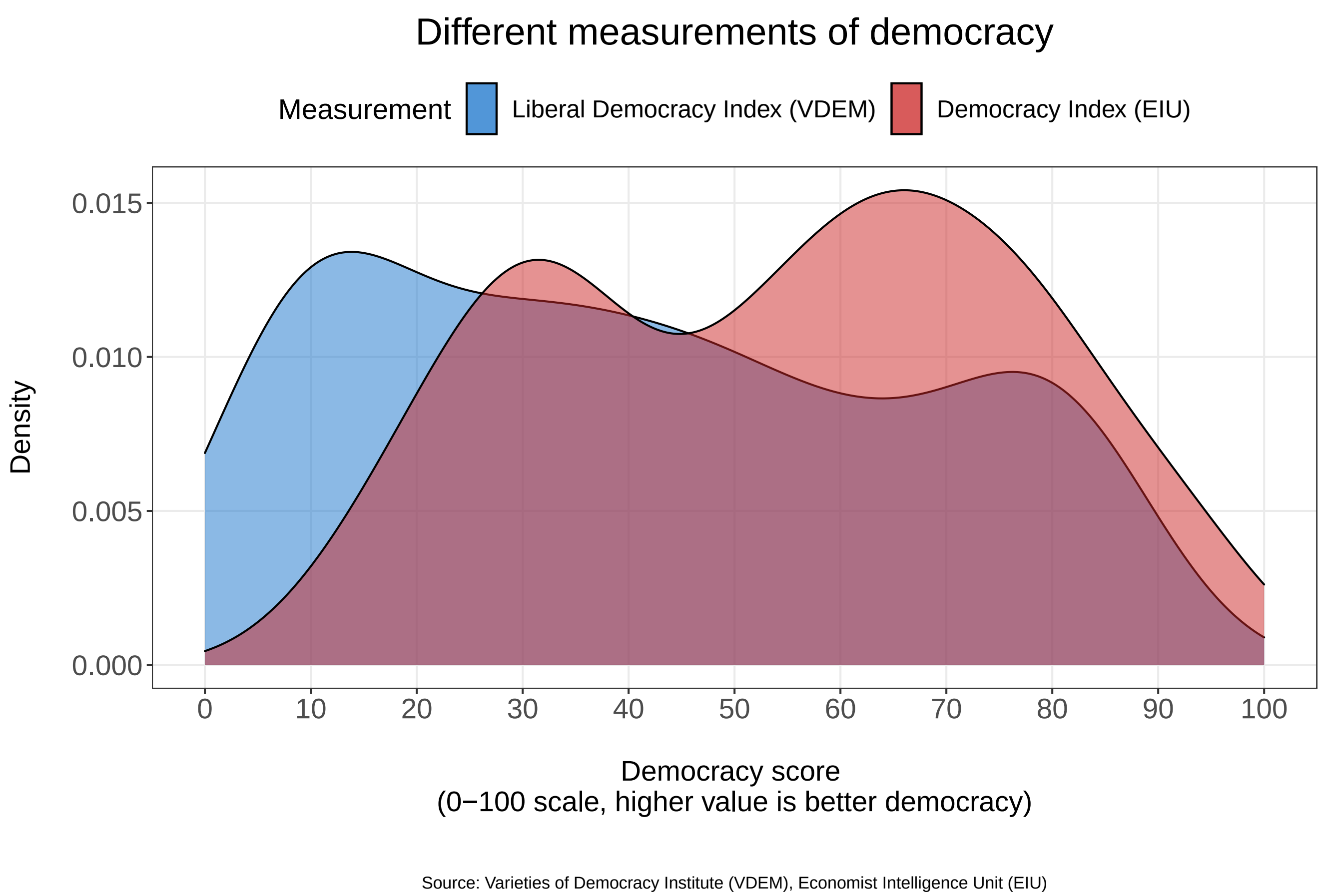

```
ggsave("density_equality.png", width = 10, height = 7)
```

```
## Warning: Removed 22 rows containing non-finite values (stat_density).
## Removed 33 rows containing non-finite values (stat_density).
```

```
# Democracy density plot

density_democracy <- ggplot(data = correlation)+
  geom_density(aes(x=libdem,
                   fill = "Liberal Democracy Index (VDEM)"),
               alpha = 0.5)+
  geom_density(aes(x = demindex,
                   fill = "Democracy Index (EIU)"),
               alpha = 0.5)+
  scale_fill_manual(values = democracy_colors,
                    labs(fill = "Measurement"))+
  scale_x_continuous(limits = c(0, 100),
                     breaks = seq(0, 100, by = 10))+
  labs(title = "Different measurements of democracy",
       caption = "\nSource: Varieties of Democracy Institute (VDEM), Economist Intelligence Unit (
       x = "\nDemocracy score \n(0-100 scale, higher value is better democracy)\n",
       y = "\nDensity\n") +
  theme_bw()+
```

```
  theme(legend.position = "top",    #Legend on top
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 15),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 13),
        legend.key.height = unit(1, 'cm'),
        panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5,
                                  size = 20,
                                  margin=margin(20,0,10,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 12,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5)
        )

density_democracy
```

```
## Warning: Removed 21 rows containing non-finite values (stat_density).

## Warning: Removed 29 rows containing non-finite values (stat_density).
```



```r
ggsave("density_democracy.png", width = 10, height = 7)
```

```
## Warning: Removed 21 rows containing non-finite values (stat_density).
## Removed 29 rows containing non-finite values (stat_density).
```

What becomes clear is that the Liberal Democracy Index has a much stricter definition of democracy. The Democracy Index is much more lenient in terms of assigning countries a higher democracy score, particularly a value between 60 and 70. Looking at economic equality, I observe that the Gini Index before tax rate countries much lower on the economic equality inequality ladder (remember that a score of 100 represents perfect inequality). This is to be expected, given that governments use the tax system to distribute recourses to secure stability and alleviate social grievances, both in democracies and autocracies. What's more, there is a noticeable "hump" in the Gini Index before tax, around a value of 45, in contrast to the smooth bell shaped form of the Gini index after tax. It's the same story in the Liberal Democracy Index. These findings will clearly have implications for the hypothesis testing. Among other things, difference of means test is more imprecise when applied to a variable which don't follow a normal distribution.

Correlation matrix

It will be very useful to know how the variables are related to each other. I use a correlation matrix to get a sense of the strength of the relationship between economic equality, democracy, and gdp per capita. I expect to see strong correlations between all the variables.

```r
# Making a matrix data frame from the merged data set

matrix <- merged_data1 %>%
  filter(time == "2019") %>%    #Year 2019
  select(c(1, 5:7, 9:10)) %>%   #only numerical main variables
  drop_na()   #Removing NAs

matrix <- data.frame(matrix, row.names = 1)   #Countries as row names

matrix_pearson <- round(cor(matrix, method = "pearson"),   #Using Pearson's r
                        10)   #Rounding up to two desimals

head(matrix_pearson[, 1:5])   #Double-checking
```

```
##                gini_after_tax gini_before_tax  gdpcapita  resources     libdem
## gini_after_tax      1.0000000       0.8046687  0.4275004 -0.0845124  0.1893361
## gini_before_tax     0.8046687       1.0000000  0.5755113 -0.2435446  0.4482718
## gdpcapita           0.4275004       0.5755113  1.0000000 -0.2364688  0.6590703
## resources          -0.0845124      -0.2435446 -0.2364688  1.0000000 -0.4021432
## libdem              0.1893361       0.4482718  0.6590703 -0.4021432  1.0000000
```

```r
# Correlation matrix plot

plot <- ggcorrplot(matrix_pearson,
           hc.order = TRUE,   #Ordering by hierarchical cluster analysis
           lab = TRUE,   #Adding correlation coefficient to the plot
```

```
           show.diag = FALSE,    #Hiding diagonal
           outline.color = "black")+
  scale_fill_gradient(high = "dodgerblue3",
                      low = "grey90",
                      name = expression(paste("Pearson's r, ", rho))
                      ) +
  scale_x_discrete(labels = c("Gini Index after tax",   #Renaming variables
                              "Gini Index before tax",
                              "GDP per capita",
                              "Liberal Democracy Index",
                              "Democracy Index")) +
  scale_y_discrete(labels = c("Gini Index after tax",
                              "Gini Index before tax",
                              "GDP per capita",
                              "Liberal Democracy Index",
                              "Democracy Index"))+
  labs(title = "Correlation matrix",
       caption = "Source: Economist Intelligence Unit, Varieties of Democracy Institute, World Ine
  theme(plot.title = element_text(size = 15,    #Space between title and top
                                  hjust = 0.5,    #Title in middle
                                  margin=margin(20,0,0,0)),
        plot.caption = element_text(hjust = 0.5)
        )    #Space in margin
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```

```
# Showing and saving plot

plot
```


```r
ggsave("matrix.png", width = 10, height = 7)
```

There is indeed correlation between many, but not all, of the variables. As expected, the different measurements of democracy, and GDP per capita, correlate with each other. But very unexpectedly, this is not true for the Gini Indexes and the democracy indexes. One exception is the Gini Index before tax, which correlates very strongly with GDP per capita. This makes intuitive sense, as higher income, and greater productivity, is often associated with lower equality overall.

5 Analysing data: hypotheses 1 and 2

In part five I finally begin analysing the data. I use both linear regression analysis and difference of means analysis, to test hypotheses 1 and 2 from multiple angles. I provide visualizations of the results.

Linear regression

Regression models

Testing hypothesis 1. Looking for correlation between economic equality and democracy. Using GDP per capita, socialism and natural resources as as control variables. To eliminate the confusing results which arises from the different units of measurement, I standardize the coefficents from the independent variables.

```r
# Standardized regression data

correlation_standardized <- correlation %>%
  select(5:11, 15) %>%
  scale() %>%
  as.data.frame()

correlation_not_standardized <- correlation %>%
  select(5:11, 15) %>%
  as.data.frame()

# Different regression models

model0 <- lm(libdem ~ gini_before_tax+gdpcapita+resources+socialism,   #Using OLS regression
             data = correlation_not_standardized,
             na.action = "na.exclude")

model1 <- lm(libdem ~ gini_before_tax+gdpcapita+resources+socialism,   #Using OLS regression
             data = correlation_standardized,
             na.action = "na.exclude")

model2 <- lm(libdem ~ gini_after_tax+gdpcapita+resources+socialism,   #Using OLS regression
             data = correlation_standardized,
             na.action = "na.exclude")

model3 <- lm(demindex ~ gini_before_tax+gdpcapita+resources+socialism,   #Using OLS regression
             data = correlation_standardized,
             na.action = "na.exclude")

model4 <- lm(demindex ~ gini_after_tax+gdpcapita+resources+socialism,   #Using OLS regression
             data = correlation_standardized,
             na.action = "na.exclude")

stargazer(model1, model2, model3, model4,  #Including all the models
          out = "stargazer2.txt",
          type = "text",   #Printing text edition
          report = "vc*p",   #Showing p-value in paranthesis
          title = "Results from standardized  regression analysis",
          dep.var.labels = c("Liberal Democracy Index",   #Dependentt variabels
                             "Democracy Index"),
          covariate.labels = c("Gini Index before tax [1]",   #Independent variables
                               "Gini Index after tax [1]",
                               "GDP per capita [2]",
                               "Natural resources [3]",
                               "Socialism [4]"),
          column.labels = c("Before tax", "After tax", #Clarifaction
                            "Before tax", "After tax"),
          notes = c("",
                    "[1] 0 = perfect inequality, 100 = perfect equality",
                    "[2] In constant 2017 international dollars",
```

```r
                    "[3] Natural resources rents as percentage of GDP",
                    "[4] Socialist regime in power for at least 60 years"),
          notes.align = "l",
          notes.label = "",
          df = FALSE)   #Hiding degrees of freedom
```

```
##
## Results from standardized regression analysis
## ============================================================================
##                                                Dependent variable:
##                            ------------------------------------------------
##                            Liberal Democracy Index      Democracy Index
##                             Before tax   After tax  Before tax   After tax
##                                (1)          (2)         (3)         (4)
## ----------------------------------------------------------------------------
## Gini Index before tax [1]   0.273***                 0.246***
##                             p = 0.0001              p = 0.0002
##
## Gini Index after tax [1]                 -0.041                  -0.038
##                                         p = 0.522               p = 0.516
##
## GDP per capita [2]          0.319***    0.596***    0.342***    0.606***
##                            p = 0.00001  p = 0.000  p = 0.00000  p = 0.000
##
## Natural resources [3]      -0.247***   -0.285***   -0.282***   -0.318***
##                            p = 0.00005 p = 0.0002  p = 0.00001 p = 0.00001
##
## Socialism [4]              -0.209***   -0.158***   -0.221***   -0.169***
##                             p = 0.001   p = 0.007  p = 0.0002   p = 0.002
##
## Constant                     0.060      0.128**      0.067      0.114**
##                             p = 0.305   p = 0.025   p = 0.229   p = 0.029
##
## ----------------------------------------------------------------------------
## Observations                  158          147        155         142
## R2                           0.456        0.516      0.508       0.591
## Adjusted R2                  0.442        0.503      0.495       0.579
## Residual Std. Error          0.729        0.675      0.687       0.609
## F Statistic                32.124***    37.875***  38.683***   49.561***
## ============================================================================
## *p<0.1; **p<0.05; ***p<0.01
##
## [1] 0 = perfect inequality, 100 = perfect equality
## [2] In constant 2017 international dollars
## [3] Natural resources rents as percentage of GDP
## [4] Socialist regime in power for at least 60 years
```

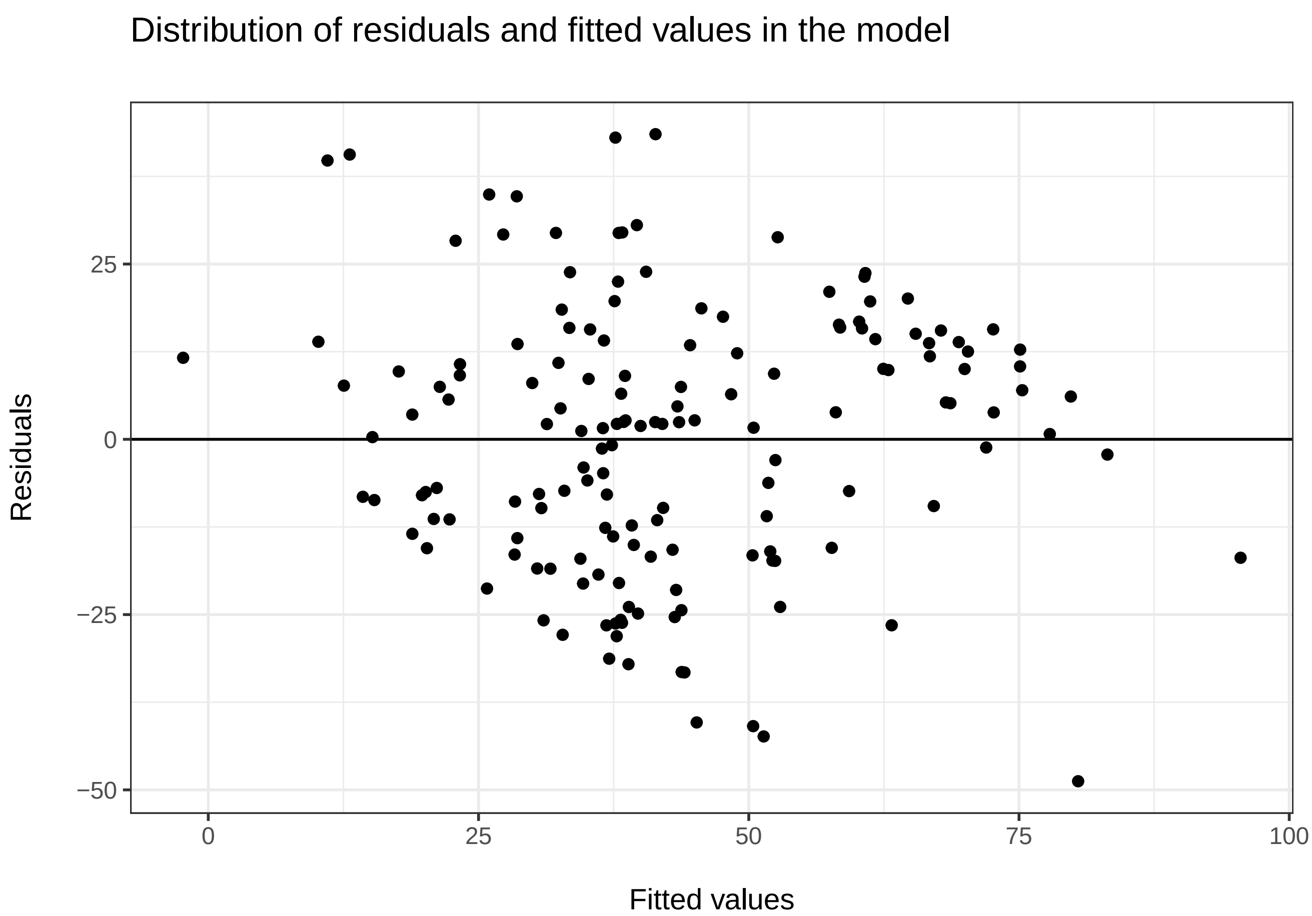
The regression analysis has some interesting results. In all the models, GDP per capita has a very statically significant impact on democracy. This is in line with the classic modernization theory in the social science literature, which states that countries are likelier to democratize as they become richer. Socialism has a noticeable negative impact on democracy, a testimony to the self-delusional "people's republic". The natural resources variable also has a harmful influence on democracy, a phenomena know as known as the paradox of plenty, or the resource curse.

This is all well and good, but I am primarily interested in how the variables measuring economic equality fared in the regression analysis. The picture which emerges, is that economic equality has a statistically significant effects on democracy, when measured before taxation. Both models 1 and 3 pass the significance threshold of 5% ($P < 0.05$). Not only is Gini Index before tax statistically significant, but it also has the second biggest effect size of all the independent variables, even surpassing natural resources. In other words; whether or not a country is economically equal or not, has a bigger say on the quality of its democracy, than whether the country is dependent or not on natural resources. Taken alone, these results clearly approve of H2. However, this interpretation is weakened by the fact that models 2 and 4 produce very statistically insignificant results ($P = 0.522$ and $P = 0.516$ respectively). Economic equality seems to have no effect on democracy when the redistributive effects of the tax system is taken into account. This suggests that even authoritarian countries redistribute income to maintain some sorts of social stability.

Model diagnostics

As model 1 included the most observations, and most easily achieved statistical significance, it is worth taking a closer look at. The model makes predictions on the values of the dependent variable in the observations, based on the values of the independent variables, in this case the Gini Index, GDP per capita and Socialism. These predictions are called fitted values. The difference between the fitted values and the observed values are called residuals. I will use diagnostic tools to test whether the assumptions made by the linear regression model are met or not. To do so, I examine the distribution of residuals and fitted values in the model. Results are reported in the non_standardizedd coefficient format.

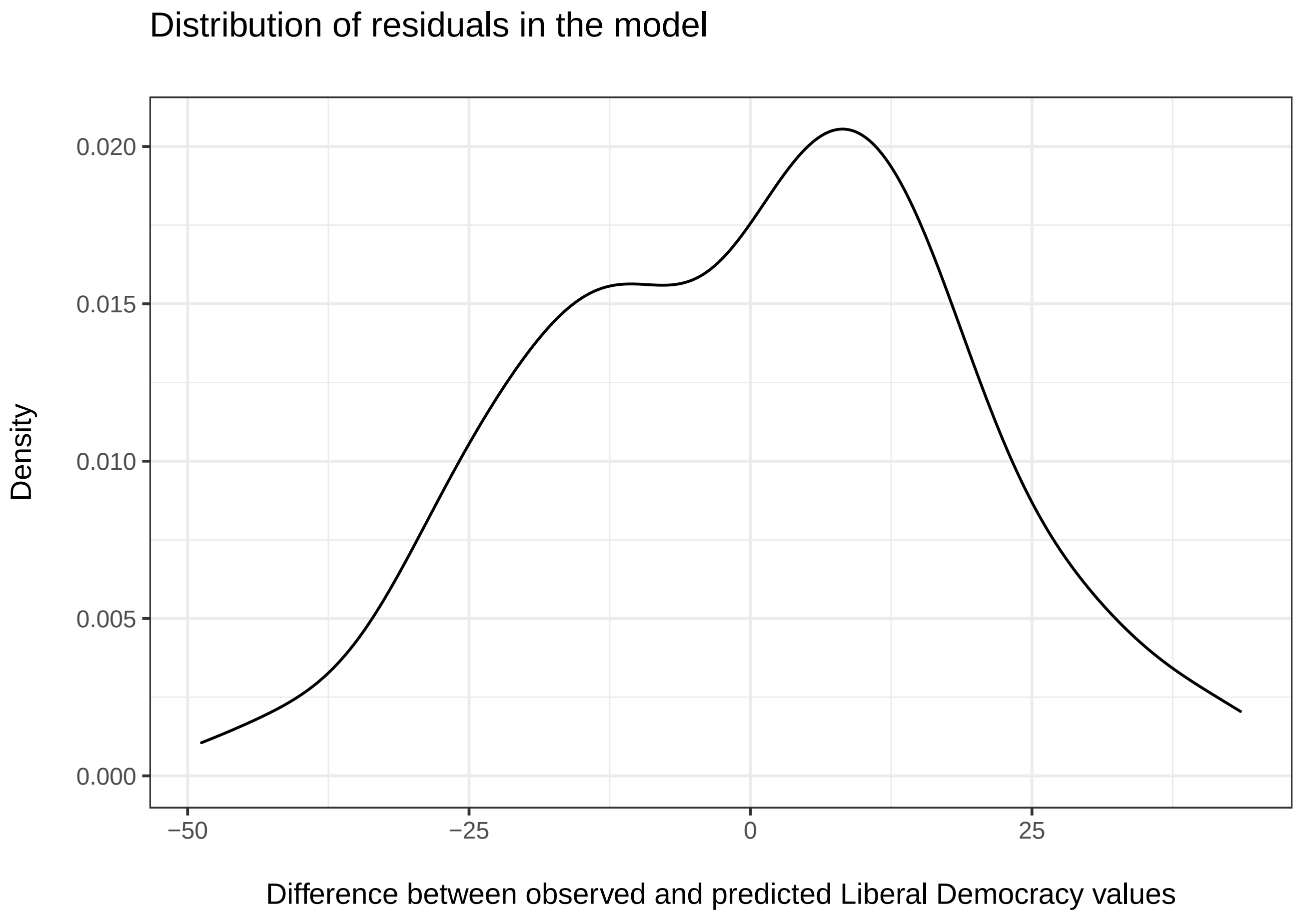
```r
# Model diagnostics data set

correlation_diagnostics <- correlation %>%
  mutate(modelfitted = fitted(model0),
         modelresid = resid(model0),
         modelresid_abs = abs(modelresid),
         modelcook = cooks.distance(model0),
         row_names = row.names(correlation)
         )

# Residuals and fitted values

ggplot(correlation_diagnostics, aes(x = modelfitted, y = modelresid)) +
  geom_point(stat = "identity")+
  geom_hline(yintercept = 0)+
  labs(x = "\nFitted values\n",
       y = "\nResiduals\n",
       title = "Distribution of residuals and fitted values in the model\n")+
  theme_bw()
```



The graph shows the relationship between the fitted values and the residuals, and should ideally appear like a cloud, which will indicate a lack of bias in the data. This is unfortunately not the case, as the distribution of the observations takes shape more like a bulge. When the model predicts a democracy score of above 60 and below 25, it is generally on point. But when it predicts a democracy score of around 30-50, it is much less accurate. The model fails to explain the high democracy scores of some Latin American countries, like Costa Rica and Singapore. It also fails to explain the particularly low democracy scores of gulf states like UAE and Qatar. Still, the predictions are not widely of the mark, not bad for a purely economic model. The variation in democracy in African and European countries are for the most part accurately captured, the fitted value of Kenya and is only 0.8 points off, for instance.

```
# Distribution of residuals in the model

ggplot(correlation_diagnostics, aes(x = modelresid)) +
  geom_density(kernel = "gaussian")+
  labs(x = "\nDifference between observed and predicted Liberal Democracy values\n",
       y = "\nDensity\n",
       title = "Distribution of residuals in the model\n")+
  theme_bw()
```

```
## Warning: Removed 36 rows containing non-finite values (stat_density).
```

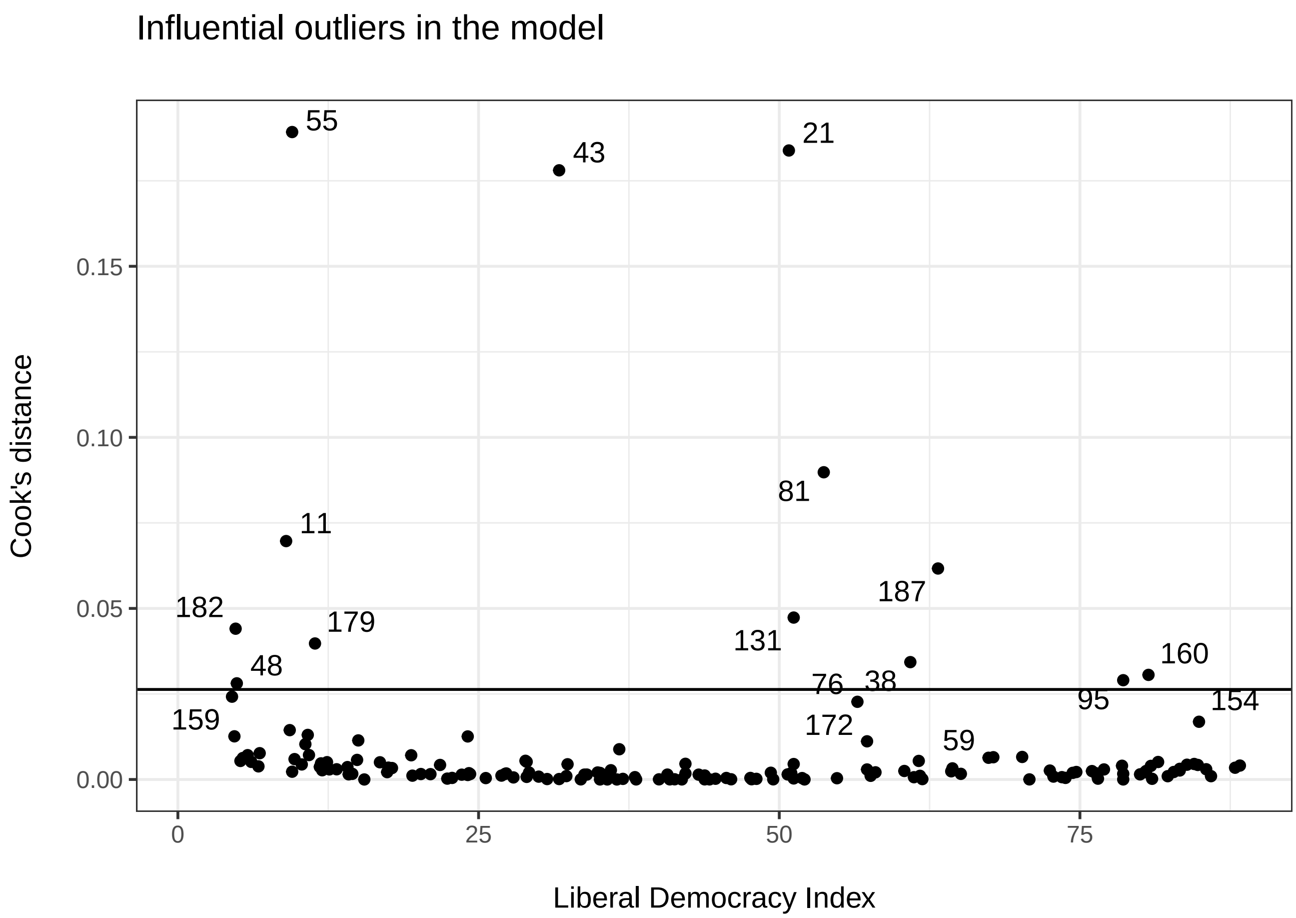


The plot shows the distribution of the residuals. It looks fairly bell-shaped, around the residual 0, suggesting that model1 has done a good job of capturing the patterns in the dataset. Even though model only has a moderately positive linear association (R2 = 0.456), this promosing results shows that it at least warrants further investigation.

```
# Identifying outliers with influence - Cook's D

N = sum (!is.na(cooks.distance(model1)))   #Number of observations
k = length(coefficients(model1)-1)   #Number of variables subtracted by 1
limit.cook = (4/(N-k-1))   #Limit value

ggplot(correlation_diagnostics,
       aes(x = libdem, y = modelcook, label = row_names)) +
  geom_point()+
  geom_text_repel()+
  geom_hline(yintercept = limit.cook)+
  labs(x = "\nLiberal Democracy Index\n",
       y = "\nCook's distance\n",
       title = "Influential outliers in the model\n")+
  theme_bw()
```

```
## Warning: Removed 36 rows containing missing values (geom_point).

## Warning: Removed 36 rows containing missing values (geom_text_repel).
```

```
## Warning: ggrepel: 140 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
correlation_diagnostics %>%
  select(row_names, country, modelcook) %>%
  arrange(-modelcook)
```

```
## # A tibble: 194 x 3
##    row_names country              modelcook
##    <chr>     <chr>                    <dbl>
##  1 55        Qatar                   0.189
##  2 21        Timor-Leste             0.184
##  3 43        Singapore               0.178
##  4 81        Mongolia                0.0898
##  5 11        United Arab Emirates    0.0697
##  6 187       Armenia                 0.0617
##  7 131       Georgia                 0.0473
##  8 182       Bahrain                 0.0441
##  9 179       Belarus                 0.0398
## 10 38        South Africa            0.0343
## # ... with 184 more rows
```

```
correlation_diagnostics %>%
  select(country, modelresid, modelresid_abs) %>%
  arrange(-modelresid_abs)
```

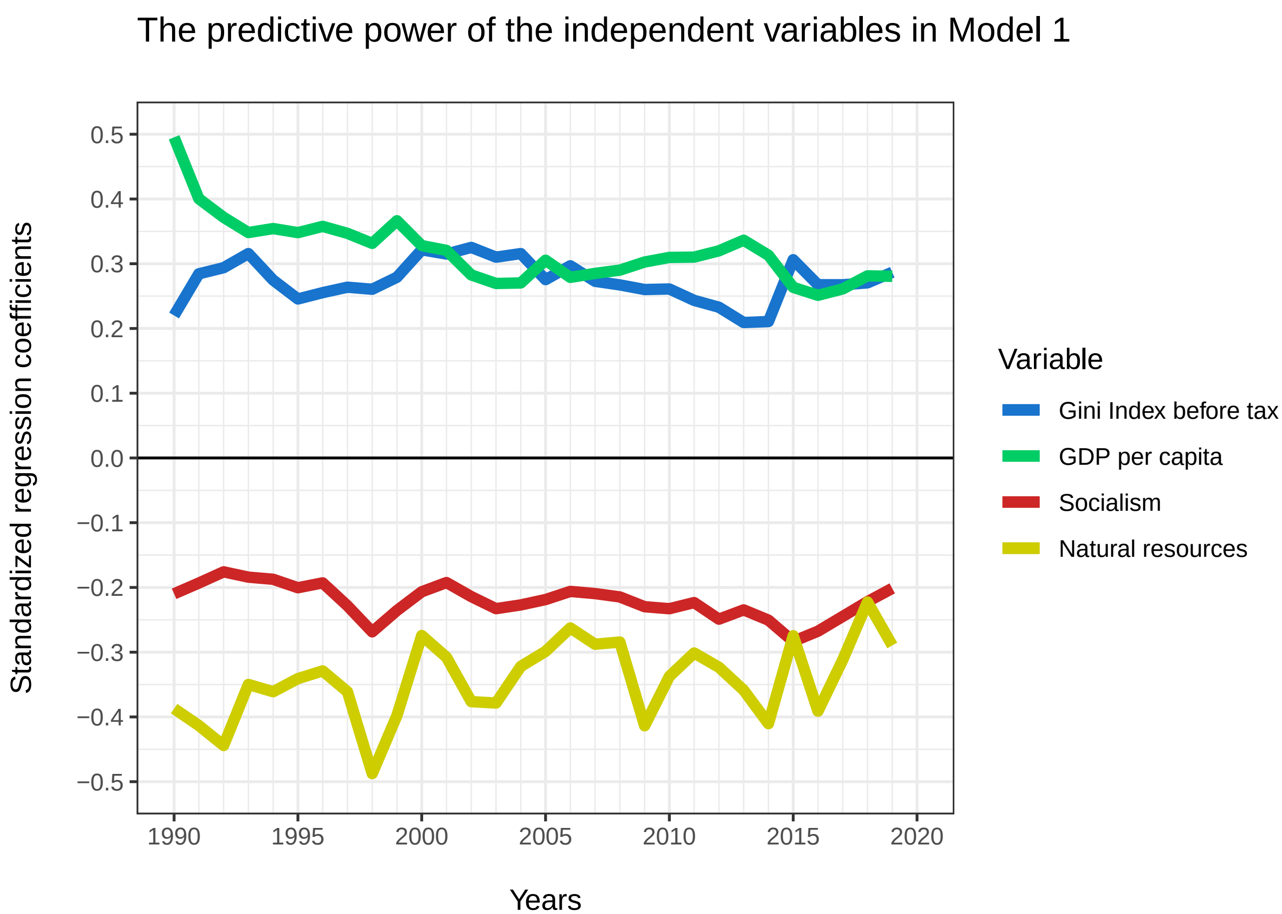
```
## # A tibble: 194 x 3
##    country              modelresid modelresid_abs
##    <chr>                     <dbl>          <dbl>
##  1 Singapore                 -48.8           48.8
##  2 Costa Rica                 43.5           43.5
##  3 Chile                      43.0           43.0
##  4 United Arab Emirates      -42.4           42.4
##  5 Qatar                     -40.9           40.9
##  6 Mongolia                   40.6           40.6
##  7 Bahrain                   -40.4           40.4
##  8 Timor-Leste                39.8           39.8
##  9 South Africa               34.9           34.9
## 10 Armenia                    34.7           34.7
## # ... with 184 more rows
```

The plot shows Cook's distance, an estimate of the influence of observations when performing a least-squares regression analysis. We see that Qatar, Timor-Leste and Singapore are influential outliers, which is not a huge surprise. They are all very small countries, with unusual kind of economies. There is only a tiny difference between countries which are influential outliers, and countries which are absolute outliers. Most notable is Georgia, a country with bigger influence on the model, than what is expected from its deviation from the fitted value. All in all, the data seems to meet the assumptions made by the linear regression model.

This positive result could be a lucky accident, a product of favourable conditions which were in place the year the data was collected. It will be very interesting to see if the beta coefficients and the coefficient of determination vary a lot by which year the regression analysis is conducted in. I will therefore check if the pattern holds over time, by looking at the different models over several years,

```
# Regression models data

regression_data <-
  merged_data1 %>%
  filter(!(str_detect(time, "past|present"))) %>%
  mutate(time = as.numeric(time)) %>%
  filter(time >= 1990) %>%
  select(time, gdpcapita, socialism, resources, gini_before_tax, gini_after_tax, demindex, libdem)
  scale() %>%
  as.data.frame()

# A list of lm objects with a common linear model for different years

regression1_coefficients <- data.frame((summary(
  lme4::lmList(libdem ~ gini_before_tax+gdpcapita+resources+socialism|time,
               data = regression_data,
               na.action = "na.exclude")))$coefficients)
```

```r
# Regression coefficients over time

regression1_coefficients$years <- (1990:2021)
regression1_coefficients$gini_before_tax <- regression1_coefficients$Estimate.gini_before_tax
regression1_coefficients$gdpcapita <- regression1_coefficients$Estimate.gdpcapita
regression1_coefficients$socialism <- regression1_coefficients$Estimate.socialism
regression1_coefficients$resources <- regression1_coefficients$Estimate.resources

# Regression coefficients colors

coefficient_colors <- c("Gini Index before tax" = "dodgerblue3",
                        "GDP per capita" = "springgreen3",
                        "Socialism" = "firebrick3",
                        "Natural resources" = "yellow3")

# Plot

ggplot(regression1_coefficients, aes(x = years))+
  geom_line(aes(y = gini_before_tax, color = "Gini Index before tax"), size = 2) +
  geom_line(aes(y = gdpcapita, color = "GDP per capita"), size = 2) +
  geom_line(aes(y = socialism, color = "Socialism"), size = 2)+
  geom_line(aes(y = resources, color = "Natural resources"), size = 2)+
  geom_hline(yintercept = 0) +
  scale_color_manual(values = coefficient_colors, labs(fill = "Variable"))+
  scale_x_continuous(limits = c(1990, 2020),
                     breaks = seq(1990, 2020, by = 5),
                     minor_breaks = seq(1990, 2020, by = 1))+
  scale_y_continuous(limits = c(-0.5, 0.5),
                     breaks = seq(-0.5, 0.5, by = 0.1))+
  labs(x = "\nYears\n",
       y = "\nStandardized regression coefficients\n",
       title = "The predictive power of the independent variables in Model 1\n")+
  theme_bw()
```

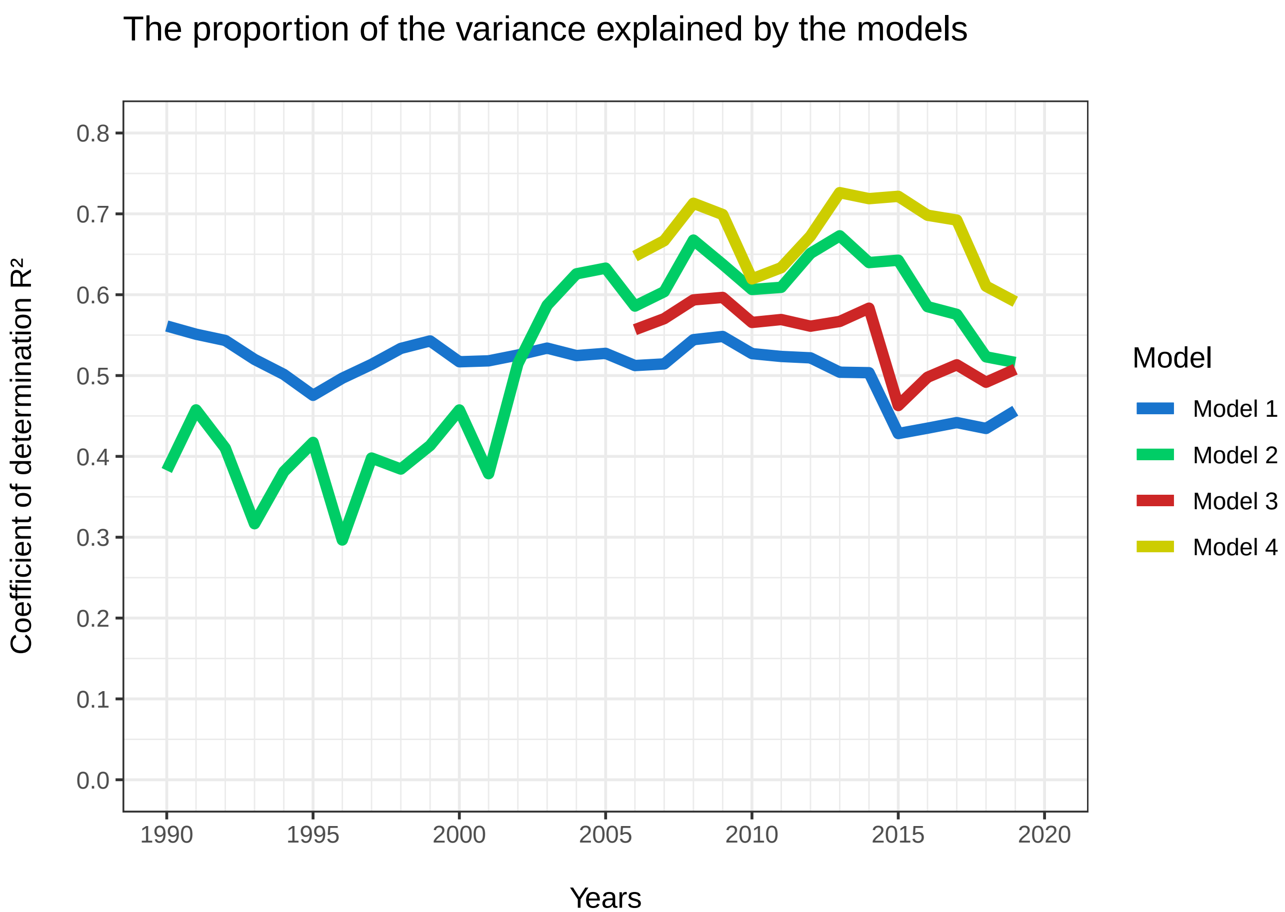


The explanation power of the different variables show some variation over time. Since the year 1990, GDP per capita has decreasingly been a reliable indicator for predicting democracy in a country. However, it is still a strong predictor. On the other hand, Gini Index before tax has increased its explanation power, now on a historic high point not seen since 2006. This may have something to do with the recent third wave of autocratization, which is particularly prevalent in middle-income and low-income countries. The global slowdown in the growth of developing countries since 2010, may also have something to with this recent trend.

```
# Several lists of lm objects with a common linear model for different years

regression1_r_squared <- data.frame((summary(
  lme4::lmList(libdem ~ gini_before_tax+gdpcapita+resources+socialism|time,
               data = regression_data,
               na.action = "na.exclude")))$r.squared)

regression1_r_squared$years <- (1990:2021)
regression1_r_squared$model1_r2 <- regression1_r_squared$X.summary.lme4..lmList.libdem...gini_befo


regression2_r_squared <- data.frame((summary(
  lme4::lmList(libdem ~ gini_after_tax+gdpcapita+resources+socialism|time,
               data = regression_data,
               na.action = "na.exclude")))$r.squared)
```

```
regression2_r_squared$years <- (1990:2021)
regression2_r_squared$model2_r2 <- regression2_r_squared$X.summary.lme4..lmList.libdem...gini_afte


regression3_r_squared <- data.frame((summary(
  lme4::lmList(demindex ~ gini_before_tax+gdpcapita+resources+socialism|time,
               data = regression_data,
               na.action = "na.exclude")))$r.squared)

regression3_r_squared$years <- (1990:2021)
regression3_r_squared$model3_r2 <- regression3_r_squared$X.summary.lme4..lmList.demindex...gini_be


regression4_r_squared <- data.frame((summary(
  lme4::lmList(demindex ~ gini_after_tax+gdpcapita+resources+socialism|time,
               data = regression_data,
               na.action = "na.exclude")))$r.squared)

regression4_r_squared$years <- (1990:2021)
regression4_r_squared$model4_r2 <- regression4_r_squared$X.summary.lme4..lmList.demindex...gini_af

# R squared over time

regression_r_squared <- regression1_r_squared %>%
  left_join(regression2_r_squared, by = "years") %>%
  left_join(regression3_r_squared, by = "years") %>%
  left_join(regression4_r_squared, by = "years")

# R squared colors

regression_colors <- c("Model 1" = "dodgerblue3",
                       "Model 2" = "springgreen3",
                       "Model 3" = "firebrick3",
                       "Model 4" = "yellow3")

# Plot

ggplot(regression_r_squared, aes(x = years))+
  geom_line(aes(y = model1_r2, color = "Model 1"), size = 2) +
  geom_line(aes(y = model2_r2, color = "Model 2"), size = 2) +
  geom_line(aes(y = model3_r2, color = "Model 3"), size = 2)+
  geom_line(aes(y = model4_r2, color = "Model 4"), size = 2)+
  scale_color_manual(values = regression_colors, labs(fill = "Model"))+
  scale_x_continuous(limits = c(1990, 2020),
                     breaks = seq(1990, 2020, by = 5),
                     minor_breaks = seq(1990, 2020, by = 1))+
  scale_y_continuous(limits = c(0, 0.8),
                     breaks = seq(0, 0.8, by = 0.1))+
  labs(x = "\nYears\n",
       y = "\nCoefficient of determination R²\n",
```

```
    title = "The proportion of the variance explained by the models\n")+
  theme_bw()
```



The proportion of the variance explained by the models show significant variation over time. Models 2 and 4 (measuring equality after tax) consistently explain the greater share of the variance, while models 1 and 3 (measuring equality before tax) are more consistent. These economic models predicted the quality of democracy in different countries, most accurately in the period between 2008 and 2014. After the financial crisis, and the advent of the third wave of autocratization, all models lost a sizeable amount of their prediction power in the late 2010s. Recently, models 2 and 4 have continued to lose relevance, while models 1 and 4 have stabilized at a moderate level. What gives? It is very conspicuous that the predictive powers of the economic models, fall at the exact same time as more countries than ever are experiencing autocratization (see chapter 7). One interpretation is that the recent wave of falling democracy around the world, is driven by something else than just pure economic factors. Non-tangible causes, such as cultural ideas, political norms, and self-expression values, could have a hand in this. These factors are beyond the scope of this analysis but may very well be decisive for a country's democracy.

Visualizing regression

Regression plot of model 1

```r
# Mean values dividing graph into quadrants

mean(correlation$gini_before_tax, na.rm = TRUE)
```

```
## [1] 43.85812
```

```r
mean(correlation$libdem, na.rm = TRUE)
```

```
## [1] 41.05607
```

```r
# The number of countries included

sum(correlation$gini_before_tax & correlation$libdem, na.rm = TRUE) #169
```

```
## [1] 169
```

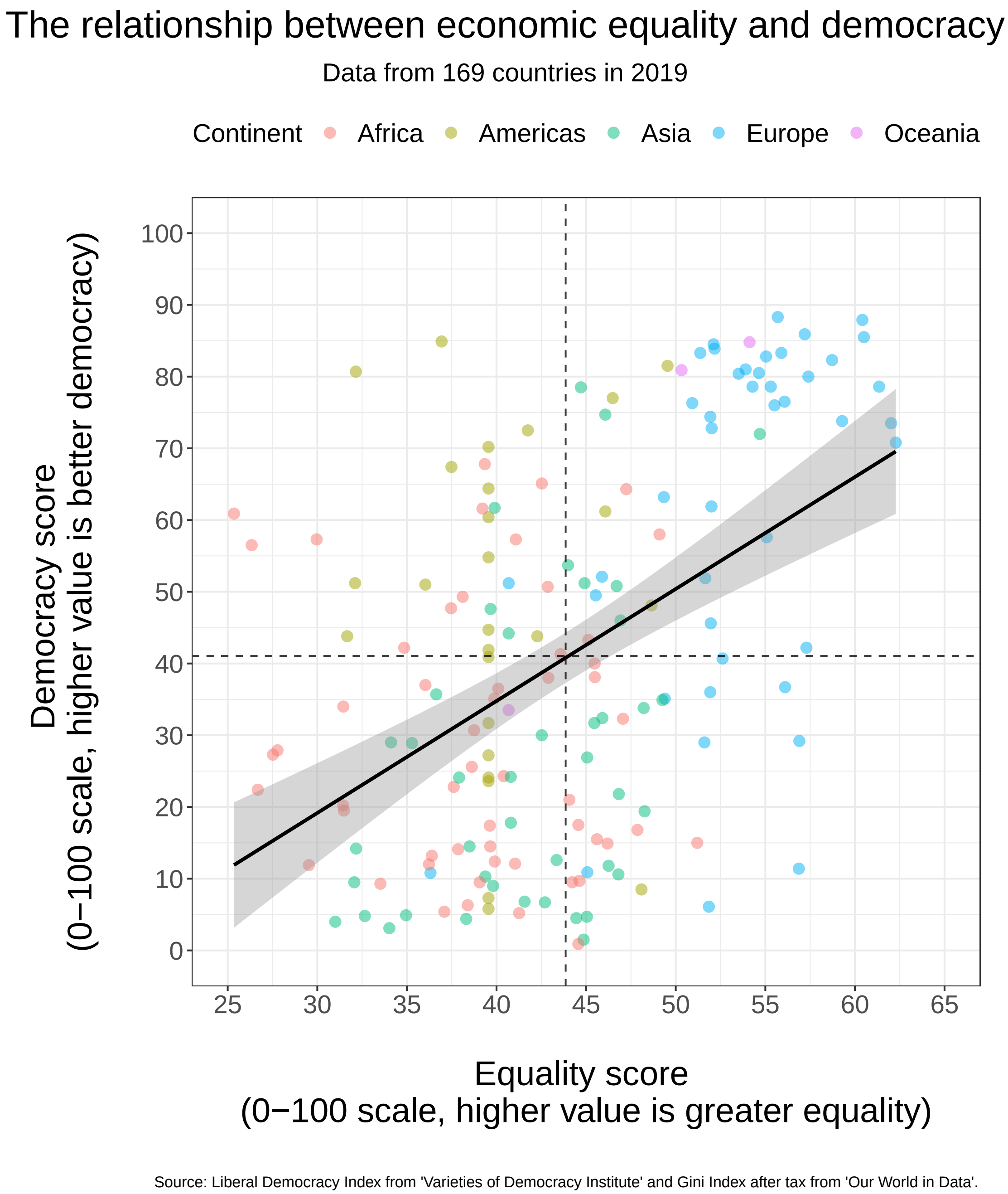
```r
# Making a plot for GINI after tax

plot <- ggplot(correlation,
               aes(y = libdem, x = gini_before_tax,
           color = Continent,
           label = Continent)) +  #Turning data into 53 countries in 2008
  geom_point(size = 3,
             alpha = 0.5,
             shape = 16)+
  geom_smooth(method = "lm",   #Linear model
              color = "black") +
  scale_y_continuous(limits = c(0, 100),
                     breaks = seq(0, 100, by = 10)) +
  scale_x_continuous(limits = c(25, 65),
                     breaks = seq(25, 65, by = 5)) +
  geom_hline(yintercept = 41.05607,
             size = 0.5,
             alpha = 0.75,
             linetype = "dashed",
             color = "black") +
  geom_vline(xintercept =  43.85812,
             size = 0.5,
             alpha = 0.75,
             linetype= "dashed",
             color = "black") +
  labs(x = "\nEquality score \n(0-100 scale, higher value is greater equality)\n",
       y = "Democracy score \n(0-100 scale, higher value is better democracy)\n",
       title = "The relationship between economic equality and democracy",
       subtitle = "Data from 169 countries in 2019",
             caption = "Source: Liberal Democracy Index from 'Varieties of Democracy Institute' an
  theme_bw() +
  theme(legend.position = "top",   #Continents on top
```

```
        plot.title.position = "plot",   #Title text in middle of the graph
        aspect.ratio = 1,  #Square graph
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 20),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.key.height = unit(1.5, 'cm'),
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(20,0,0,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,0,0))
        )

# Showing and saving plot

plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 25 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 25 rows containing missing values (geom_point).
```




```r
ggsave("regression_plot.png", width = 10, height = 10)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 25 rows containing non-finite values (stat_smooth).
## Removed 25 rows containing missing values (geom_point).
```

Faceted regression plots

```r
# Making a facet dataframe

correlation_regression <- correlation %>%
  rename(`Gini Index after tax (OWID)` = gini_after_tax,
         `Gini Index before tax (WID)` = gini_before_tax,
         `Liberal Democracy Index (VDEM)` = libdem,
         `Democracy Index (EIU)` = demindex) %>%
  pivot_longer(cols = c(10,11),
               names_to = "libdem_demindex",
               values_to = "democracy") %>%
  pivot_longer(cols = 5:6,
              names_to = "before_after_tax",
              values_to = "gini")

# Mean values dividing graph into quadrants

mean(correlation_regression$gini, na.rm=TRUE)
```

```
## [1] 52.61249
```

```r
mean(correlation_regression$democracy, na.rm=TRUE)
```

```
## [1] 47.59704
```

```r
# The number of countries included

sum(correlation$gini_before_tax & correlation$libdem, na.rm = TRUE)
```

```
## [1] 169
```

```r
sum(correlation$gini_before_tax & correlation$demindex, na.rm = TRUE) #164
```

```
## [1] 164
```

```r
sum(correlation$gini_after_tax & correlation$libdem, na.rm = TRUE) #154
```

```
## [1] 154
```

```r
sum(correlation$gini_after_tax & correlation$demindex, na.rm = TRUE) #147
```

```
## [1] 147
```

```
# Making a facet plot

facet_regression <- ggplot(correlation_regression,
               aes(x = gini, y = democracy))+
  geom_point(aes(colour = Continent))+   #Colored by continent
  scale_y_continuous(limits = c(0, 100),
                     breaks = seq(0, 100, by = 20)) +
  geom_smooth(color = "black")+
  geom_hline(yintercept = 47.32128,
             size = 0.4,
             alpha = 0.75,
             linetype = "dashed",
             color = "black") +
  geom_vline(xintercept = 47.59704,
             size = 0.4,
             alpha = 0.75,
             linetype= "dashed",
             color = "black") +
  facet_grid(factor(before_after_tax,
                    levels = c("Gini Index before tax (WID)",
                               "Gini Index after tax (OWID)")
                    ) ~ libdem_demindex
             ) +
  labs(x = "\nCountry equality score \n(0-100 scale, higher value is greater equality)\n",
       y = "Country democracy score \n(0-100 scale, higher value is better democracy)\n",
       title = "The relationship between economic equality and democracy",
       subtitle = "Using 2 different measurements of both equality and democracy, in around 160 co
       caption = "Source: Economist Intelligence Unit (EIU), Varieties of Democracy Institute (VDE
  guides(colour = guide_legend(override.aes = list(size=5)))+
  theme_bw() +
  theme(legend.position = "top",   #Continents on top
        plot.title.position = "plot",   #Title text in middle of the graph
        aspect.ratio = 1,  #Square graph
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 20),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.title = element_blank(),
        legend.text = element_text(size = 15),
        legend.key.height = unit(1, 'cm'),
        plot.title = element_text(hjust = 0,
                                  size = 25,
                                  margin=margin(20,0,0,0)),
        plot.subtitle = element_text(hjust = 0,
                                     size = 15,
                                     margin=margin(10,0,0,0))
        )


facet_regression
```

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

## Warning: Removed 142 rows containing non-finite values (stat_smooth).

## Warning: Removed 142 rows containing missing values (geom_point).
```


```r
# Showing and saving plot
ggsave("facet_regression.png", width = 11, height = 11)
```

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

```
## Warning: Removed 142 rows containing non-finite values (stat_smooth).
## Removed 142 rows containing missing values (geom_point).
```

As we can see on the graph, there seems to be no clear relationship between democracy, and Gini Index after tax. By comparison, the relationship seems more straightforward between democracy and Gini Index before tax. This is in line with the regression analysis. We can also pinpoint a cluster of equal and democratic European countries, which seems to be the main driver of the relationship. This hints at the possibility of Europe being an historical outlier, in the world of equality and democacy.

Difference of means

Preliminary tests

I perform preliminary tests to check if the conditions to perform a two-samples t-test hold up. First I check if there is homogeneity of variance

```r
# Performing the F-test to see if there is homogeneity of variance

var.test(demindex ~ equality_before_tax,   # p = 0.86
         data = correlation)
```

```
##
##  F test to compare two variances
##
## data:  demindex by equality_before_tax
## F = 0.86149, num df = 92, denom df = 70, p-value = 0.4998
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5495473 1.3324253
## sample estimates:
## ratio of variances
##          0.8614865
```

```r
var.test(demindex ~ equality_after_tax,   # p = 0.81
         data = correlation)
```

```
##
##  F test to compare two variances
##
## data:  demindex by equality_after_tax
## F = 0.80748, num df = 86, denom df = 59, p-value = 0.3621
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.498151 1.281656
## sample estimates:
## ratio of variances
##          0.8074754
```

```r
var.test(libdem ~ equality_before_tax,    # p = 0.69
         data = correlation)
```

```
##
##  F test to compare two variances
##
## data:  libdem by equality_before_tax
## F = 0.68743, num df = 95, denom df = 72, p-value = 0.08711
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.441492 1.056283
## sample estimates:
## ratio of variances
##           0.687431
```

```r
var.test(libdem ~ equality_after_tax,    # p = 0.66
         data = correlation)
```

```
##
##  F test to compare two variances
##
## data:  libdem by equality_after_tax
## F = 0.66479, num df = 91, denom df = 61, p-value = 0.07642
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4142365 1.0443796
## sample estimates:
## ratio of variances
##          0.6647908
```

```r
# Performing the Shapiro-Wilk test to see if the data is normally distrubuted

shapiro.test(correlation$demindex)    # p = 0.00046
```

```
##
##  Shapiro-Wilk normality test
##
## data:  correlation$demindex
## W = 0.96612, p-value = 0.0004629
```

```r
shapiro.test(correlation$libdem)    # p = 0.00000057
```

```
##
##  Shapiro-Wilk normality test
##
## data:  correlation$libdem
## W = 0.93596, p-value = 5.656e-07
```

```r
shapiro.test(correlation$gini_before_tax)   # p = 0.048
```

```
##
##  Shapiro-Wilk normality test
##
## data:  correlation$gini_before_tax
## W = 0.98413, p-value = 0.04751
```

```r
shapiro.test(correlation$gini_after_tax)   # p = 0.00053
```

```
##
##  Shapiro-Wilk normality test
##
## data:  correlation$gini_after_tax
## W = 0.96591, p-value = 0.0005304
```

Although the data pass the F-test with ease, they fail the Shapiro-Wilk test. None of the p-values from the latter test, have a value greater than 0.05, implying that the distributions of the data are significantly different from a normal distribution. In other words, I cannot assume normality in the data. This rules out the use of a classic two-samples t-test. I will instead use a more sophisticated unpaired two-samples test, called the Wilcoxon rank-sum test (also known as the Mann–Whitney U test), suited for the kind of non-parametric data I am dealing with.

Unpaired two-samples test

```r
# Performing the Wilcoxon rank-sum test

wilcox.test(demindex ~ equality_before_tax,   # p = 0.0000002
            data = correlation, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  demindex by equality_before_tax
## W = 1735, p-value = 2.022e-07
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(demindex ~ equality_after_tax,   # p = 0.0045
            data = correlation, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  demindex by equality_after_tax
## W = 1889.5, p-value = 0.004544
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(libdem ~ equality_before_tax,   # p = 0.000000077
            data = correlation, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  libdem by equality_before_tax
## W = 1810, p-value = 7.668e-08
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(libdem ~ equality_after_tax,   # p = 0.0063
            data = correlation, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  libdem by equality_after_tax
## W = 2110, p-value = 0.006298
## alternative hypothesis: true location shift is not equal to 0
```

All the p-values are below at least 0.0045, which is less than the significance level of 0.05. This indicates a lower than 5% risk of concluding that a difference exists, when there is no actual difference. I therefore reject the null hypothesis and conclude that the average democracy score of equal countries, are significantly different from unequal countries. There is much to indicate that equal countries are more democratic than unequal countries.

Visualizing difference of means

Visualizing the findings from the difference of means analysis, to see if equal countries have a significantly higher or lower democracy score, compared to unequal countries.

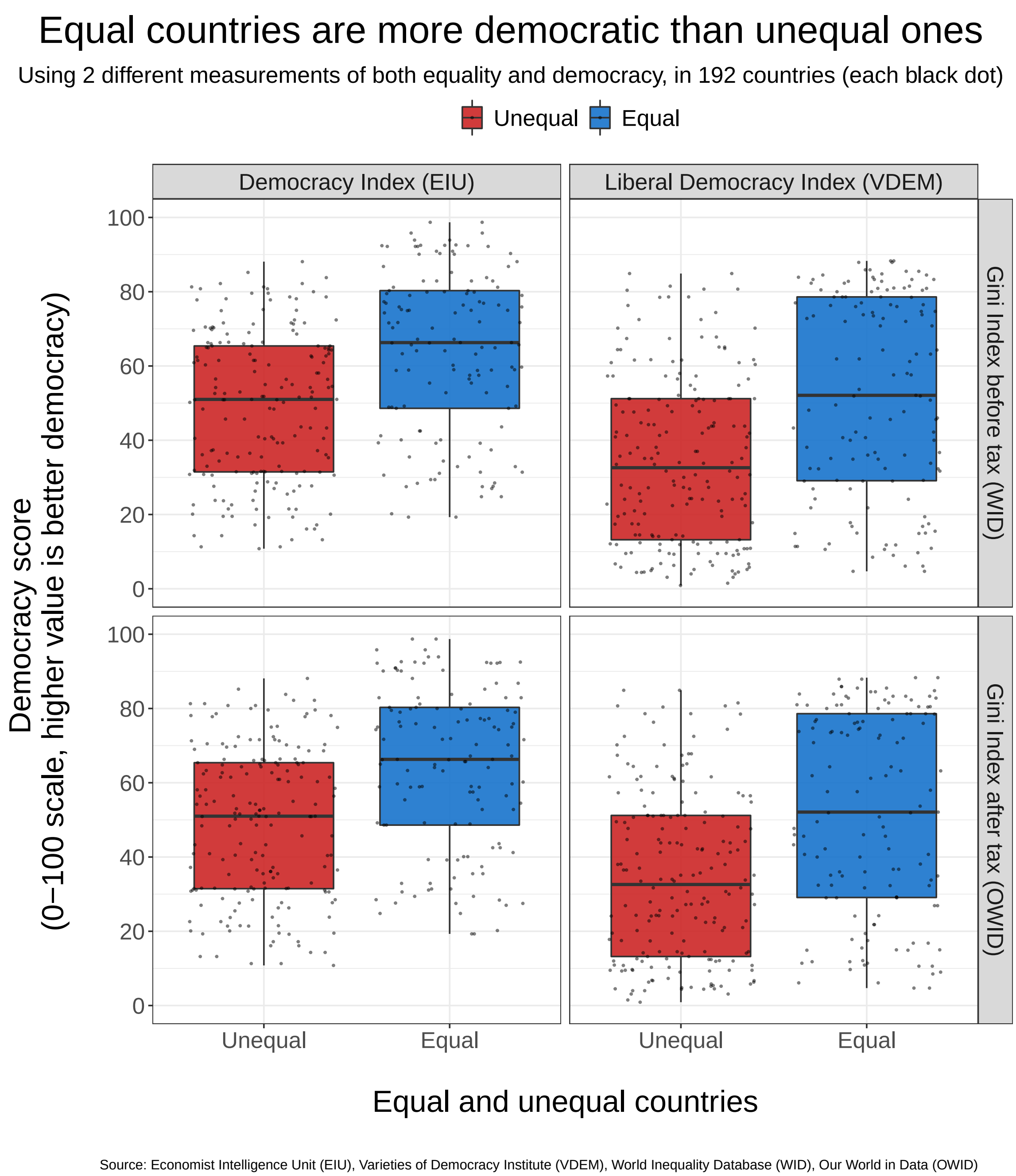
Boxplots

```r
# Boxplot colors

boxplot_colors <- c("Unequal" = "firebrick3",
                    "Equal" = "dodgerblue3")

# Data of means

correlation_means <- correlation %>%
  rename(`Gini Index after tax (OWID)` = gini_after_tax,
         `Gini Index before tax (WID)` = gini_before_tax,
         `Liberal Democracy Index (VDEM)` = libdem,
         `Democracy Index (EIU)` = demindex,
         `Before tax` = equality_before_tax,
         `After tax` = equality_after_tax) %>%
```

```
  pivot_longer(cols = 10:11,
               names_to = "libdem_demindex",
               values_to = "democracy") %>%
  pivot_longer(cols = 5:6,
              names_to = "before_after_tax",
              values_to = "gini") %>%
  pivot_longer(cols = 8:9,
              names_to = "labels_before_after",
              values_to = "Categories")


# Faceting the box plot

facet_boxplot <- ggplot(correlation_means %>%
                     filter(!is.na(Categories)),
               aes(x = democracy,
                   y = Categories,
                   fill = Categories)) +
  geom_boxplot(alpha = 0.9)+
  geom_jitter(color="black", size=0.4,    #Showing observations as dots
              alpha=0.5) +
  coord_flip() +  #Making it vertical
  scale_x_continuous(limits = c(0, 100), breaks = seq(0, 100, by = 20)) +
  scale_fill_manual(values = boxplot_colors,
                    labs(fill = "Countries by equality status"))+
  facet_grid(factor(before_after_tax,
                    levels = c("Gini Index before tax (WID)",
                               "Gini Index after tax (OWID)")
                    ) ~ libdem_demindex
             ) +
  labs(x = "Democracy score \n(0-100 scale, higher value is better democracy)\n",
       y = "\nEqual and unequal countries\n",
       title = "Equal countries are more democratic than unequal ones",
       subtitle = "Using 2 different measurements of both equality and democracy, in 192 countries
       caption = "Source: Economist Intelligence Unit (EIU), Varieties of Democracy Institute (VDE
  guides(colour = guide_legend(override.aes = list(size=5)))+
  theme_bw() +
  theme(legend.position = "top",   #Continents on top
        plot.title.position = "plot",   #Title text in middle of the graph
        aspect.ratio = 1,  #Square graph
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 20),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.title = element_blank(),
        legend.text = element_text(size = 15),
        legend.key.height = unit(1, 'cm'),
        plot.title = element_text(hjust = 0.5,
                                  size = 25,
                                  margin=margin(20,0,0,0)),
```

```
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,0,0))
        )

# Showing and saving plot

facet_boxplot
```

```
## Warning: Removed 64 rows containing non-finite values (stat_boxplot).

## Warning: Removed 64 rows containing missing values (geom_point).
```



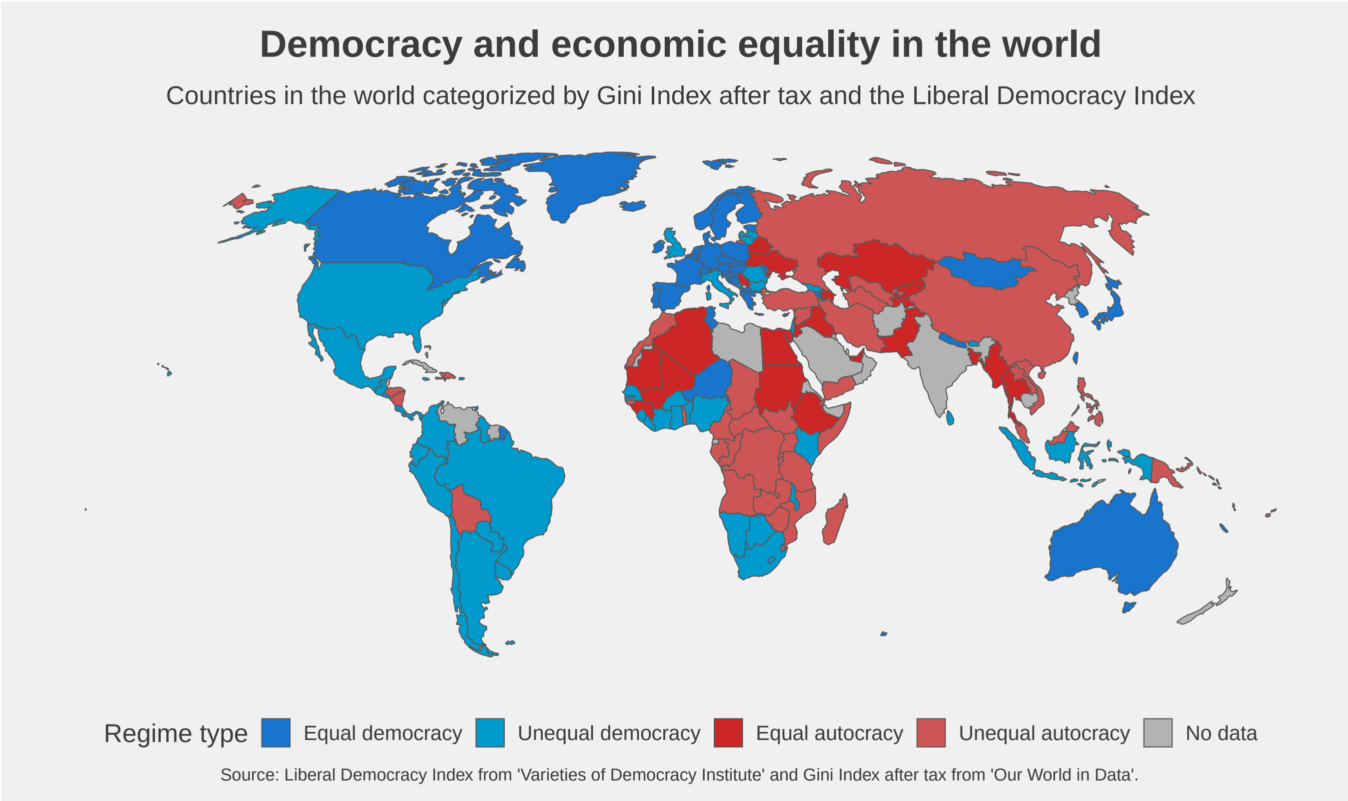
```r
ggsave("facet_boxplot.png", width = 11, height = 11)
```

```
## Warning: Removed 64 rows containing non-finite values (stat_boxplot).
## Removed 64 rows containing missing values (geom_point).
```

Equal countries appear to have a significantly higher democracy score than unequal countries. The findings strengthen hypothesis H2. Equal countries tend to be democracies, or at least more democratic, than unequal countries.

Map of covered data

```r
# Map colors

map2_colors <- c("Equal democracy" = "dodgerblue3",
                 "Unequal democracy"= "deepskyblue3",
                 "Equal autocracy"= "firebrick3",
                 "Unequal autocracy" = "indianred3",
                 "No data" = "gray70")

# Map making

map2 <- ggplot(data = map_data1) +
  geom_sf(aes(fill = regime), lwd = 0.3) +   #Adjusting border
  scale_fill_manual(values = map2_colors,
                    na.value = "gray70",
                    labs(fill = "Regime type"))+
  coord_sf(crs ="+proj=robin +lon_0=0 +x_0=0   #A better map projection
           +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")+
  labs(title = "Democracy and economic equality in the world",
       subtitle = "Countries in the world categorized by Gini Index after tax and the Liberal Demo
       caption = "Source: Liberal Democracy Index from 'Varieties of Democracy Institute' and Gini
  theme_fivethirtyeight()+   #Pleasent theme
  theme(panel.grid.major = element_blank(),   #Removing longditude and latitude
        axis.text = element_blank(),   #No unnecessary titles
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(2,0,2,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 12))


# Showing and saving map

map2
```




```r
ggsave("map2.png", width = 10, height = 7)
```

This map is a very useful visual tool for categorizing the world in terms of two simple variables: whether or not a country is a democracy/autocracy or equal/unequal. It is also useful for spotting the absent countries, where there is no data available. The world's next biggest country, India, provides no any up-to-date data of their income inequality. Several other isolationist and turbulent countries are also absent: Venezuela, Libya, Afghanistan, and Cuba. Several regional trends can be spotted on the map. Firstly, the Americas is very unequal, with Canada being the only country categorized as "equal". Secondly, Europe is very equal, regardless of whether or not the countries are democracies or autocracies. These regional patterns point to the fact that historical processes may explain a bigger role in the formation of current levels of equality and democracy, than I had previously thought of. In the Americas, the influence of settler-based societies from colonial states is an obvious explanation. In Europe, the levelling effects of socialism and the world wars may have reduced inequality significantly, compared to other world regions.

6 Analysing data: hypothisis 3

In part six I use tabular analysis to test hypothesis 3. I provide visualizations of the results.

Tabular analysis

According to H3, countries that are increasingly economically equals should experience democratization, while countries that are decreasingly equal should experience autocratization. To test the validity of this hypothesis, I will conduct a simple count of the number of countries, behaving according to the expected pattern.

Countries approving of the hypothesis is assigned the value "TRUE", while disapproving countries is assigned a "FALSE" value. In addition, I make a new variable with the four different combinations of change in equality and democracy: less equal-autocratization, less equal-democratization, more equal-autocratization, and more equal-democratization.

A question which arises is when exactly to pinpoint the starting and ending date of the analysis. I risk getting very different results depending on which years I compare the values of the Gini Index and the Liberal Democracy Index, especially for those notoriously hard define countries in the grey zone between democracy and autocracy. In addition, several financial and political crises may drastically distort the rankings of countries for some years. This bias may reduce the validity of the analysis. I landed on using the average values from the variables from 2006/2007/2008 and 2018/2019/2020 in my analysis. This has three main advantages: (1) reducing bias by using average values across two 3-year-periods, (2) including more data from countries with data from only 1 or 2 years in the 3-year-periods, and (3) avoiding a source of bias from the financial crisis in 2009 and the covid pandemic in 2020 (taken into account in the data from 2021).

```
# Liberal Democracy Index (VDEM) and GINI after tax

hypothesis <- merged_data1 %>%
  filter(time == "past") %>%  #Only the avarage data from 2006/2007/2008
  select(country,   #GDP per capita not necessary for this analysis
         id,
         gini_before_tax,
         gini_after_tax,
         libdem,
         demindex) %>%
  left_join(select(filter(merged_data1,  #Adding the avarage data from
                          time == "present"),   #2018/2019/2020 as new variable
                   country,
                   gini_before_tax,
                   gini_after_tax,
                   demindex,
                   libdem),
            by = "country") %>%
  rename(gini_past_before_tax = gini_before_tax.x,   #Renaming so it makes sense
         gini_present_before_tax = gini_before_tax.y,
         gini_past_after_tax = gini_after_tax.x,
         gini_present_after_tax = gini_after_tax.y,
         libdem_past = libdem.x,
         libdem_present = libdem.y,
         demindex_past = demindex.x,
         demindex_present = demindex.y) %>%
  mutate(gini_before_tax_change = gini_present_before_tax - gini_past_before_tax,
         gini_after_tax_change = gini_present_after_tax - gini_past_after_tax,
         libdem_change = libdem_present - libdem_past,
         demindex_change = demindex_present - demindex_past,
         dir_gini_before = ifelse(gini_before_tax_change > 0,
                                  "More equal",
                                  "Less equal"),
         dir_gini_after = ifelse(gini_after_tax_change > 0,
                                 "More equal",
```

```
                                 "Less equal"),
         dir_libdem = ifelse(libdem_change > 0,    #Character variable for change
                             "Democratization",
                             "Autocratization"),
         dir_demindex = ifelse(demindex_change > 0,
                               "Democratization",
                               "Autocratization"),
         gini_before_tax_condition = ifelse(gini_before_tax_change < 0,
                                            FALSE,
                                            TRUE),
         gini_after_tax_condition = ifelse(gini_after_tax_change < 0,
                                           FALSE,
                                           TRUE),
         libdem_condition = ifelse(libdem_change > 0,
                                   TRUE,
                                   FALSE),
         demindex_condition = ifelse(demindex_change > 0,
                                     TRUE,
                                     FALSE),
         libdem_before_tax = ifelse(gini_before_tax_condition ==
                                      libdem_condition,
                                    TRUE,
                                    FALSE),
         libdem_after_tax = ifelse(gini_after_tax_condition ==
                                     libdem_condition,
                                   TRUE,
                                   FALSE),
         demindex_before_tax = ifelse(gini_before_tax_condition ==
                                        demindex_condition,
                                      TRUE,
                                      FALSE),
         demindex_after_tax = ifelse(gini_after_tax_condition ==
                                       demindex_condition,
                                     TRUE,
                                     FALSE),
         combination = paste0(dir_gini_before,    #Usefult to keep track
                              '-',
                              dir_libdem)
         )

# Variables detecting missing values

hypothesis$na_detected = !(str_detect(hypothesis$combination, "NA"))

hypothesis$combination2 = case_when(hypothesis$combination ==
                                      "More equal-Autocratization" ~ 2,
                                    hypothesis$na_detected == TRUE ~ 1)


# Hypothesis testing
```

libdem_before_tax	n
FALSE	84
TRUE	84
NA	26

libdem_after_tax	n
FALSE	43
TRUE	58
NA	93

demindex_before_tax	n
FALSE	73
TRUE	87
NA	34

demindex_after_tax	n
FALSE	46
TRUE	50
NA	98

```
list1 <- hypothesis %>%
  count(libdem_before_tax)

list2 <- hypothesis %>%
  count(libdem_after_tax)

list3 <- hypothesis %>%
  count(demindex_before_tax)

list4 <- hypothesis %>%
  count(demindex_after_tax)

list5 <- (list(list1, list2, list3, list4))

kable(list5, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

```
# Adding the hypothesis testing to my merged data

merged_data2 <- merge(merged_data1, hypothesis, by = "country") %>%
  arrange(desc(time))

merged_data2 <- merged_data2[-c(389:12389), ] #Removing all years in time-variabel

merged_data2 <- merged_data2 %>%
  mutate(population_levels =
           cut(merged_data2$population/10^6,
               breaks = c(0, 20, 100, 500, 2000),
               labels = c("<20m","20m-100m","100m-500m", "500m+"))
         )
```

The results are slightly promising but provide an overall mixed picture. Three out of the four different test case (based on the different measurements of both democracy and economic equality) give a positive result. This indicates that more countries are approving the hypothesis, than disapproving of it. The most promising results come the Democracy Index and Gini Index before tax. In this case 53% of countries behaved according to the hypotheses, they are becoming either more equal and democratic, or less equal and less democratic. However, 73 countries still disapproved of the hypothesis, even in the most promising

combination	n
Less equal-Autocratization	37
Less equal-Democratization	25
More equal-Autocratization	59
More equal-Democratization	47
More equal-NA	4
NA-Democratization	4
NA-NA	18

case. Moreover, the results vary wildly in the different cases. There is a high number of missing country-observations in the Gini Index after tax, which tells me proceed with caution in making wider generalisations about the findings.

```
list6 <- hypothesis %>%
  count(combination)

kable(list6, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

On closer inspection, the close results from the test cases turns out to be because a significant number of countries (59), are becoming more equal, despite undergoing a process of autocratization. Fewer countries (25) are becoming more democratic, if they are already less economically equal.

Visualizing tabular analysis

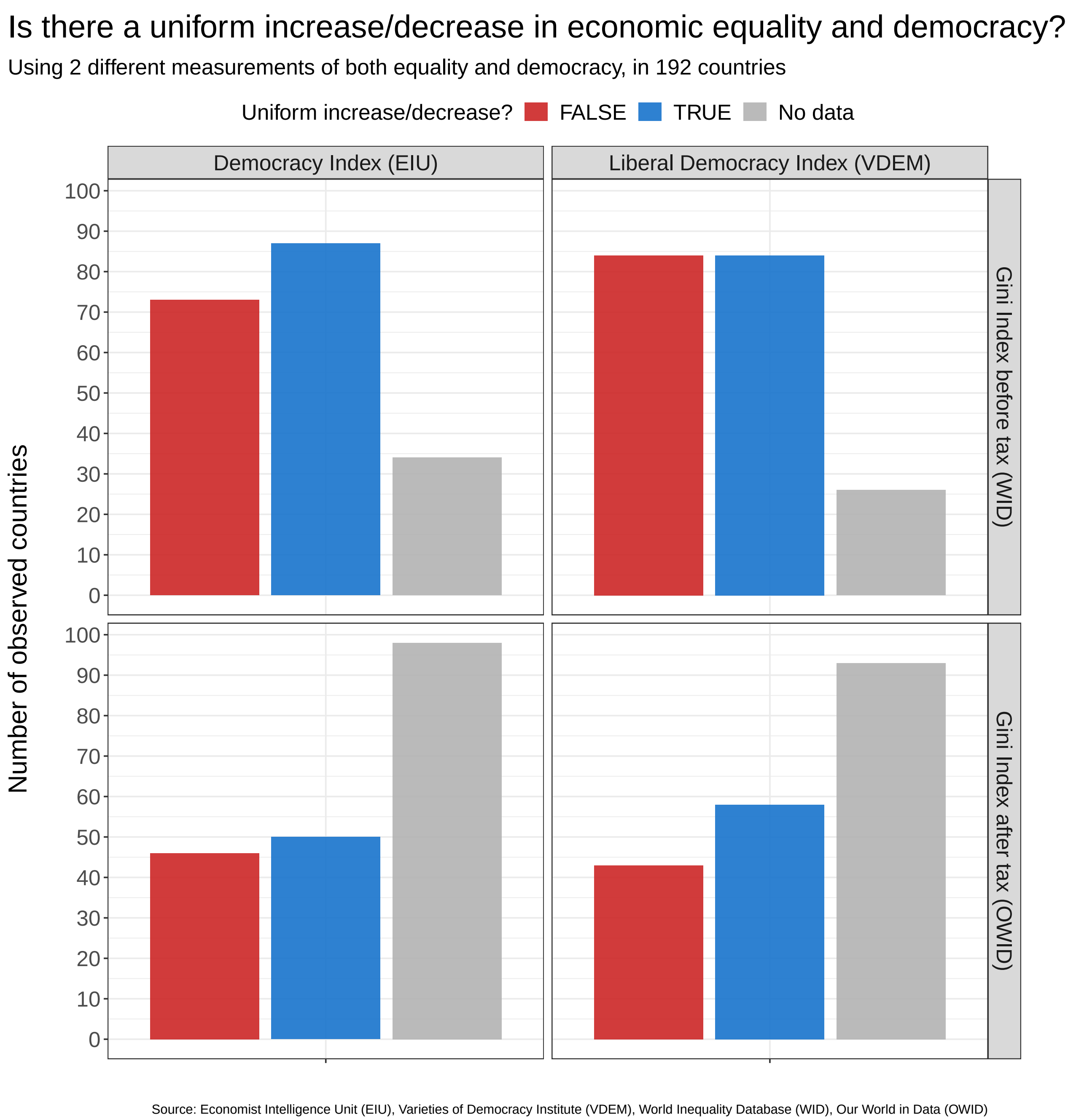
Barplot

A barplot of the above findings.

```
# Bar plot colors

bar_colors <- c("FALSE" = "firebrick3",
                "TRUE" = "dodgerblue3",
                "No data" = "gray70")

# Making a data frame from the hypothesis data set

hypothesis_barplot <- hypothesis %>%
  pivot_longer(c(libdem_before_tax,
                 libdem_after_tax,
                 demindex_before_tax,
                 demindex_after_tax),
               names_to = "measurements",
               values_to = "value") %>%
  select(measurements, value) %>%
  group_by(measurements, value) %>%
  tally()
```

```r
# Adding variables

hypothesis_barplot$libdem_demindex <-
  c(rep("Democracy Index (EIU)", 6),
    rep("Liberal Democracy Index (VDEM)", 6))

hypothesis_barplot$before_after_tax <-
  c(rep("Gini Index after tax (OWID)", 3),
    rep("Gini Index before tax (WID)", 3),
    rep("Gini Index after tax (OWID)", 3),
    rep("Gini Index before tax (WID)", 3))

# Faceting a barplot

facet_barplot <- ggplot(data = hypothesis_barplot,
                        aes(x = "",
                            y = n,
                            fill = value)) +
  geom_bar(width = 1,
           alpha = 0.9,
           stat = "identity",
           position = "dodge2")+
  scale_y_continuous(breaks = seq(0, 120, by = 10))+
  scale_fill_manual(values = bar_colors,
                    na.value = "gray70",
                    labs(fill = "Uniform increase/decrease?"))+
  facet_grid(factor(before_after_tax,
                    levels = c("Gini Index before tax (WID)",
                               "Gini Index after tax (OWID)")
                    ) ~ libdem_demindex) +
  labs(x = element_blank(),
       y = "Number of observed countries\n",
       title = "Is there a uniform increase/decrease in economic equality and democracy?",
       subtitle = "Using 2 different measurements of both equality and democracy, in 192 countries
       caption = "Source: Economist Intelligence Unit (EIU), Varieties of Democracy Institute (VDE
  theme_bw() +   #Pleasent theme
  theme(legend.position = "top",   #Continents on top
        plot.title.position = "plot",   #Title text in middle of the graph
        aspect.ratio = 1,  #Square graph
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 18),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        legend.key.height = unit(0.5, 'cm'),
        plot.title = element_text(hjust = 0,
                                  size = 22,
                                  margin=margin(20,0,0,0)),
        plot.subtitle = element_text(hjust = 0,
```

```
                                            size = 15,
                                            margin=margin(10,0,10,0))
        )

facet_barplot
```




```
ggsave("facet_barplot.png", width = 11, height = 11)
```

Planar graph

A planar graph of above findinds, which shows individual countries, instead of grouping them together into categories. I use Liberal Democracy Index and the Gini Index before tax.

```
# Finding means which will be shown as dotted lines

mean(correlation$libdem, na.rm = TRUE)
```

```
## [1] 41.05607
```

```
mean(correlation$gini_before_tax, na.rm = TRUE)
```

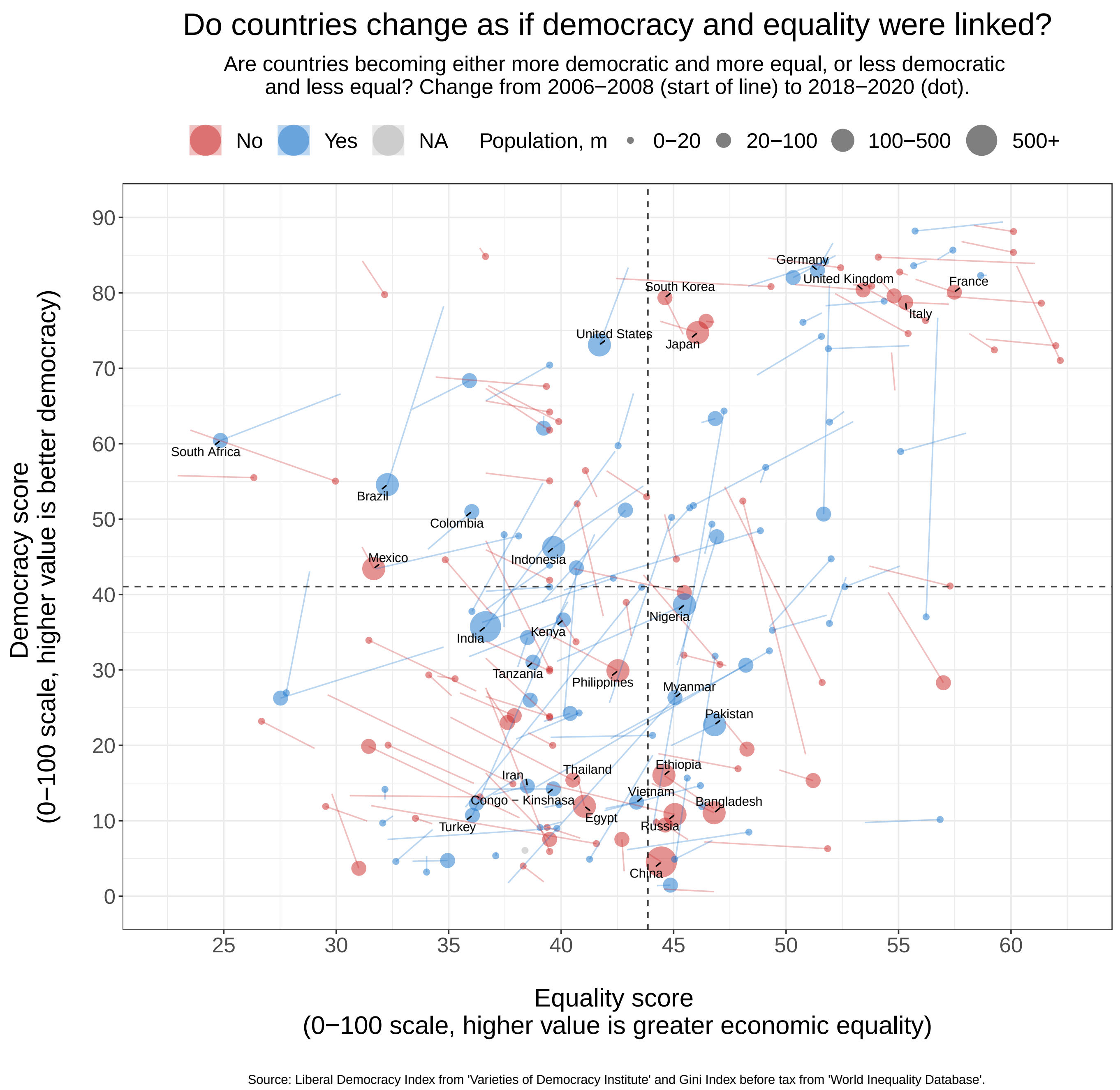
```
## [1] 43.85812
```

```
# Making a function to make life easier

country.select <- merged_data2 %>%
  filter(time == "present") %>%
  select(country)

# Planar graph of Gini Index before tax and the Liberal Democracy Index

plot <- ggplot(data = merged_data2 %>%
                 filter(time == "present") %>%
                 filter(country %in% country.select$country),
               aes(y = libdem,
                   x = gini_before_tax,
                   color = libdem_before_tax, #Coloring by the TRUE/FALSE-values
                   group = country,
                   label = country)) +
  geom_point(aes(size = as.numeric(population_levels)),   #Adjusted by pop size
             alpha = 0.5,   #Slightly transparant
             shape = 16) +   #Circles
  geom_path(alpha = 0.3,   #Slightly transparant
            data = filter(merged_data2,
                          country %in% country.select$country)) +
  geom_hline(yintercept = 41.05607,   #Avarage democracy
             size = 0.5,
             alpha = 0.75,
             linetype = "dashed",
             color = "black") +
  geom_vline(xintercept = 43.85812,   #Avarage economic equality
             size = 0.5,
             alpha = 0.75,
             linetype = "dashed",
             color = "black") +
  geom_text_repel(force = 1,
                  size = 3,
```

```
                  color = "black",   #Country labels as black
                  segment.color = "black",
                  show.legend = FALSE,   #Removing text from legend
                  min.segment.length = 0,   #Always lines to the country labels
                  data = subset(merged_data2,   #The most populous countries
                                population > 50000000 &
                                  time %in% "present")
                  ) +
  scale_x_continuous(limits = c(22.5, 62.5),
                     breaks = seq(20, 65, by = 5)) +
  scale_y_continuous(limits = c(0, 90),
                     breaks = seq(00, 90, by = 10)) +
  scale_colour_manual(values = c("firebrick3", "dodgerblue3", "gray70"),
                      na.value = "gray70",
                      labels = c("No", "Yes"),
                      name = element_blank(),
                      ) +
  scale_radius(range = c(2, 10),   #Geom point size by population
               labels = c("0-20", "20-100",  "100-500", "500+"),
               guide = guide_legend(title = "Population, m",
                                    title.hjust = 0.5)) +
 labs(x = "\nEquality score \n(0-100 scale, higher value is greater economic equality)\n",
      y = "Democracy score \n(0-100 scale, higher value is better democracy)\n",
      title = "Do countries change as if democracy and equality were linked?",
      subtitle = "Are countries becoming either more democratic and more equal, or less democratic
      caption = "Source: Liberal Democracy Index from 'Varieties of Democracy Institute' and Gini
  theme_bw()+   #Pleasent theme
  theme(legend.position = "top",   #Legends on top
        strip.text = element_text(size = 15),
        axis.title = element_text(size = 18),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 15),
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(20,0,0,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5)
        ) +
  guides(color = guide_legend(order = 1, #Color legend to the left
                              override.aes = list(size = 10)))

# Showing and saving plot

plot
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
```

```
## Warning: Removed 51 row(s) containing missing values (geom_path).
```




```r
ggsave("tabular_plot.png", width = 11, height = 11)
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
## Removed 51 row(s) containing missing values (geom_path).
```

What becomes evident when looking at the planar graph, is just how diverse the world is. Some equal and democratic countries, like Germany and France, don't seem to have changed all that much the past 15 years. On the other hand, there are countries that have experienced truly dramatic shift. India for example, has become much more authoritarian and unequal, while Nigeria is moving closer to becoming a more democratic and equal society. As the bar plot previously illustrated, there is no discernible pattern of substantially more countries behaving according to H3. All in all, more countries seem to change vertically than horizontally, indicating that it is easier to change the form of government, than the economic fabric of society.

Two proportions test

I want to know whether the proportions of democratizing countries, are the same in the two groups of more or less economically equal countries. I use a two-proportions test to compare the two observed proportions, and see if the difference between the two is statistically significant. My sample size is sufficient large, consisting of only 194 countries, to use a standard z-test.

```
hypothesis_frequencies <- data.frame(
  "More_equal" = c(47, 59, 47+59),
  "Less_equal" = c(25, 37, 25+37),
  "Total" = c(47+25, 39+57, NA),
  row.names = c("More_democratic", "Less_democratic", "Total"),
  stringsAsFactors = FALSE)

prop.test(x = c(47, 25), n = c(106, 62))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(47, 25) out of c(106, 62)
## X-squared = 0.11983, df = 1, p-value = 0.7292
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1270566  0.2073975
## sample estimates:
##    prop 1    prop 2
## 0.4433962 0.4032258
```

The p-value of the test is 0.7292, which is way more than the significance level of 0.05. This indicates a higher than 5% risk of concluding that a difference exists, when there is no actual difference. I therefore accept the null hypothesis and conclude that equal countries didn't undergo democratization significantly more often than unequal countries.

Visualizing two proportions test

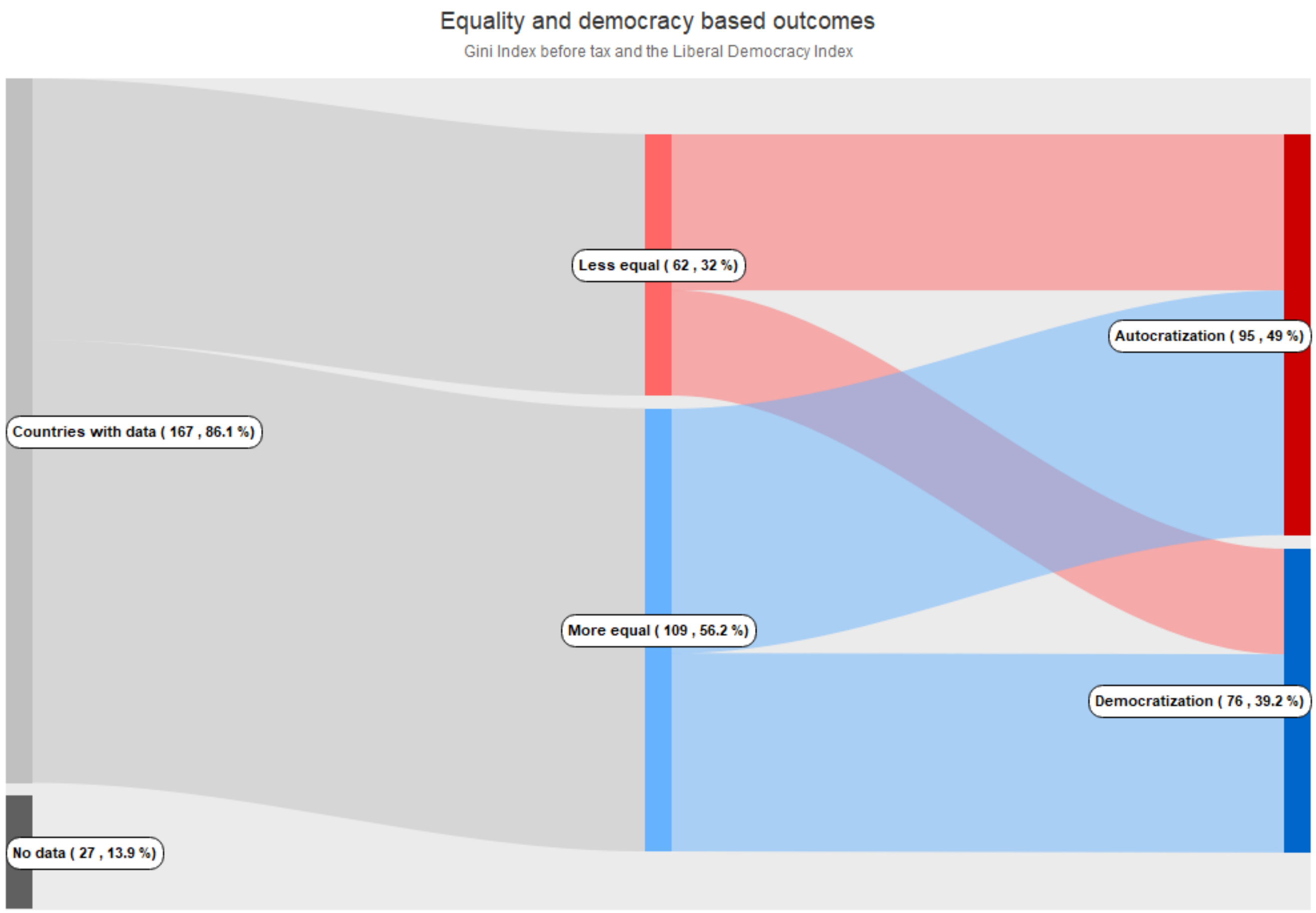
Sankey diagram

While the planar graphs shown above are visually pleasing to look at, they can be hard to pull any meaningful generalisations from. I have therefore made a sankey diagram, to sort all the countries according to their behaviour in regards to equality and democracy.

```
# Showing how many countries I have data from (coverage)

data_sankey <- correlation %>%
```

```
  select(country)

sd <- data_sankey %>%   #sd is an abbreviation for sankey data
  left_join(hypothesis, by = "country") %>%
  mutate(coverage = case_when(
    is.na(dir_gini_before) ~ "No data",
    is.na(dir_libdem) ~ "No data",
    TRUE ~ "Countries with data")
  ) %>%
  select(country, coverage, dir_gini_before, dir_libdem)

# Getting a count of gini direction and adding percentage values as a new column

sd2 <- sd %>%
  group_by(dir_gini_before) %>%
  tally() %>%
  mutate(perc = n/sum(n)) %>%
  mutate(dir_gini_new = paste(dir_gini_before, '(', n, ',', round(perc* 100,1) , '%)')) %>%
  select(-n, - perc)

sdMain <- merge (sd, sd2, by  = "dir_gini_before")

# Getting a count of democracy direction and adding percentage values as a new column

sd3 <- sd %>%
  group_by(dir_libdem) %>%
  tally() %>%
  mutate(perc = n/sum(n)) %>%
  mutate(dir_libdem_new = paste(dir_libdem, '(', n, ',', round(perc* 100,1) , '%)')) %>%
  select(-n, - perc)

sdMain <- merge (sdMain, sd3, by  = "dir_libdem")


# Getting a count of number of countries with data and adding percentage values as a new column

sd4 <- sd %>%
  group_by(coverage) %>%
  tally() %>%
  mutate(perc = n/sum(n)) %>%
  mutate(coverage_new = paste(coverage, '(', n, ',', round(perc* 100,1) , '%)')) %>%
  select(-n, - perc)

sdMain <- merge (sdMain, sd4, by  = "coverage")

# Making a final data set for the sankey diagram

sdFinal <- sdMain %>%
  select(coverage_new, dir_gini_new, dir_libdem_new)
```

```r
sdFinal[c(168:194), c(2:3)] <- NA


# Changing the colours in the nodes manually

sankey <- highchart() %>%
  hc_title(text = "Equality and democracy based outcomes") %>%
  hc_subtitle(text = "Gini Index before tax and the Liberal Democracy Index")  %>%
  hc_add_series(data = data_to_sankey(sdFinal),
                type = "sankey",
                nodes = list(list(id = "Countries with data ( 167 , 86.1 %)",
                                  color = "#C0C0C0"),
                             list(id = "No data ( 27 , 13.9 %)",
                                  color = "#606060"),
                             list(id = "Less equal ( 62 , 32 %)",
                                  color = "#FF6666"),
                             list(id = "More equal ( 109 , 56.2 %)",
                                  color = "#66B2FF"),
                             list(id = "Autocratization ( 95 , 49 %)",
                                  color = "#CC0000"),
                             list(id = "Democratization ( 76 , 39.2 %)",
                                  color = "#0066CC"))
  ) %>%
  hc_plotOptions(series = list(dataLabels = list(style = list(color = "black"),
                                                 backgroundColor = "white",
                                                 borderRadius = 10,
                                                 borderWidth = 1,
                                                 borderColor = "black",
                                                 padding = 5,
                                                 shadow = FALSE))) %>%
  hc_add_theme(hc_theme_ggplot2())   #Nice ggplot familiar theme

htmlwidgets::saveWidget(widget = sankey, file = "sankey.html") #Saving html

webshot::webshot(url = "sankey.html",   #Saving to png file on pc
        file = "sankey.png",
        vwidth = 1000,
        vheight = 700,
        delay = 3) #The delay ensures that the whole plot appears in the image
```




```
#sankey
```

The diagram makes it evident that surprisingly many countries are becoming mer economically equal, while also experiencing autocratization. The diagram also makes makes one trend depressingly clear: way more countries are experiencing autocratization than democratization. This is in line with recent research from the VDEM institutet, which suggeststs that the world is currently embroiled in a third wave of global autocratization.

Map of more equal and authoritarian countries

The class of countries becoming more equal and more authoritarian, deserves special attention. Their mere existence disapproves of my hypothesis, and goes directly against the argument that equality and democracy are inherently correlated. The first step is to identify these countries.

```
# Identifying countries becoming more egalitarian and authoritarian

hypothesis %>%
  filter(combination %in% "More equal-Autocratization") %>%
  unique()
```

```
## # A tibble: 59 x 29
##    country     id    gini_past_befor~ gini_past_after~ libdem_past demindex_past
```

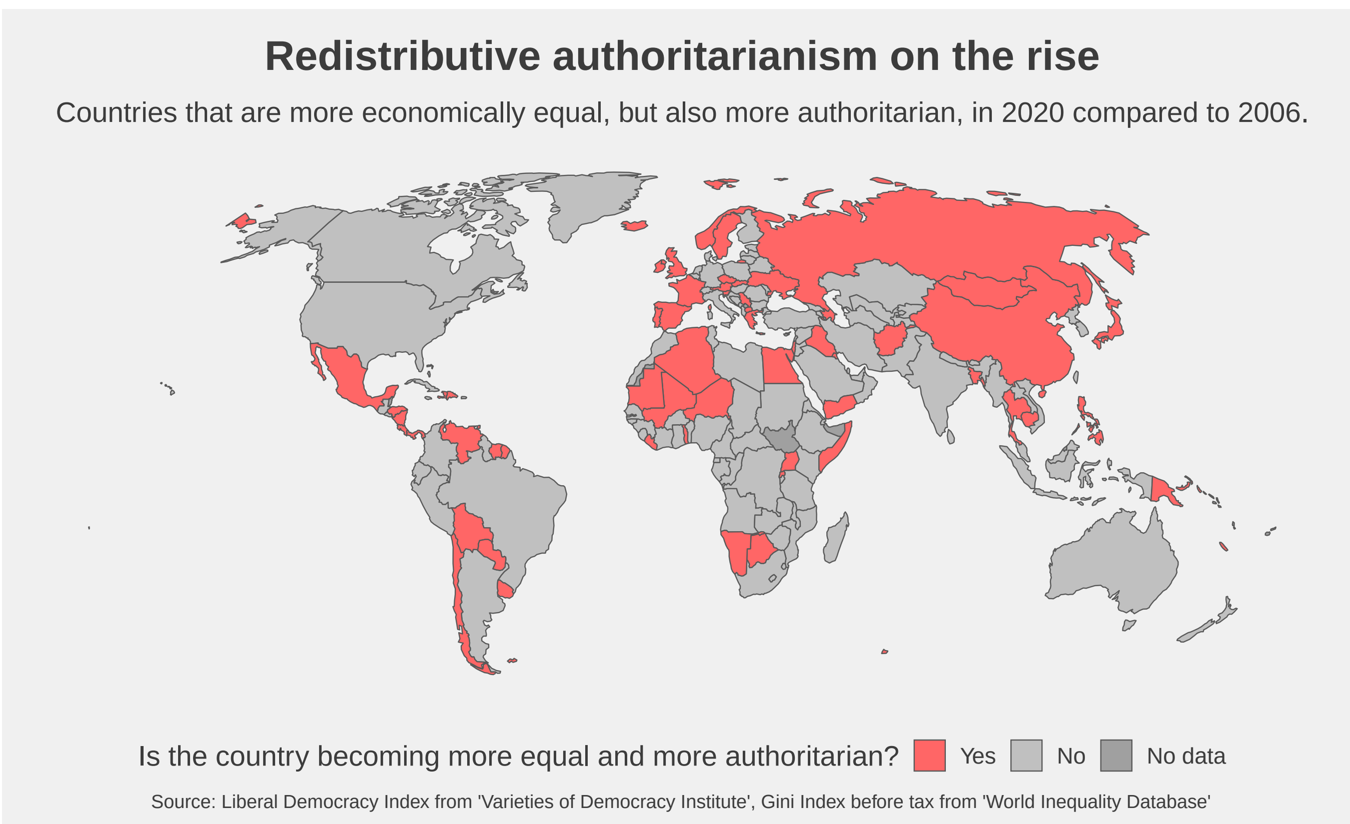
```
##    <chr>       <chr>            <dbl>            <dbl>      <dbl>         <dbl>
##   1 Yemen       YEM               29.8             NA         13.6          29.6
##   2 Venezuela   VEN               36.6             55.2       16.3          53.8
##   3 Uruguay     URY               42.4             54.2       81.9          80.2
##   4 United Kin~ GBR               50.4             65.2       81.1          81.2
##   5 Ukraine     UKR               54.5             72.2       40.3          69.4
##   6 Uganda      UGA               36.7             NA         27.1          50.8
##   7 Trinidad &~ TTO               36.6             59.7       65.7          72.0
##   8 Togo        TGO               38.5             57.8       21.7          20.9
##   9 Thailand    THA               35.1             59.4       23.7          62.4
## 10 Sweden      SWE               58.3             72.8       88.9          98.8
## # ... with 49 more rows, and 23 more variables: gini_present_before_tax <dbl>,
## #   gini_present_after_tax <dbl>, demindex_present <dbl>, libdem_present <dbl>,
## #   gini_before_tax_change <dbl>, gini_after_tax_change <dbl>,
## #   libdem_change <dbl>, demindex_change <dbl>, dir_gini_before <chr>,
## #   dir_gini_after <chr>, dir_libdem <chr>, dir_demindex <chr>,
## #   gini_before_tax_condition <lgl>, gini_after_tax_condition <lgl>,
## #   libdem_condition <lgl>, demindex_condition <lgl>, ...
```

```r
# Map data

map_data2 <- raw_map_data %>%
  select(sovereignt, geometry) %>%
  filter(sovereignt != "Antarctica") %>%
  rename(country = sovereignt) %>%
  mutate(id = countrycode(country,
                          origin = "country.name",
                          destination = "iso3c")) %>%
  left_join(hypothesis, by = "id") %>%
  mutate(combination2 = case_when(
    combination2 == 2 ~ "Yes",
    combination2 == 1 ~ "No"))
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```r
# Map colors

map3_colors <- c("Yes" = "#FF6666",
                 "No" = "#C0C0C0",
                 "No data" = "#A0A0A0")

# Map making

map3 <- ggplot(data = map_data2) +
  geom_sf(aes(fill = combination2), lwd = 0.3) +   #Adjusting border
  scale_fill_manual(values = map3_colors,
                    na.value = "#A0A0A0",
                    labs(fill = "Is the country becoming more equal and more authoritarian?"))+
  coord_sf(crs ="+proj=robin +lon_0=0 +x_0=0    #A better map projection
           +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")+
```

```
  labs(title = "Redistributive authoritarianism on the rise",
       subtitle = "Countries that are more economically equal, but also more authoritarian, in 202
       caption = "Source: Liberal Democracy Index from 'Varieties of Democracy Institute', Gini In
  theme_fivethirtyeight()+    #Pleasent theme
  theme(panel.grid.major = element_blank(),    #Removing longditude and latitude
        axis.text = element_blank(),    #No unnecessary titles
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(2,0,2,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 12))

# Showing and saving map

map3
```




```
ggsave("map3.png", width = 10, height = 7)
```

As it turns out, way more countries than expected have become more equal and authoritarian the last

20 years. This process of redistributive authoritarianism seems to apply for every continent in the world, including countries as different from another as Japan, Egypt and Chile.

7 Analysing data: hypothisis 4

In part seven I test hypothesis 4, by making an overview of how different countries have changed their levels of democracy and equality over time. I also aggregate the results for the world at large. In the end, I provide visualizations of the results.

Preparing data

```
# Liberal Democracy Index

merged_time1 <- merged_data1 %>%
  filter(!(str_detect(time, "past|present"))) %>%   #Only years
  mutate(time = as.numeric(time)) %>%   # Years are numeric
  filter(time >= 1980) %>%   #There is only data from the last 40 years
  select(country, time, libdem) %>%
  drop_na() %>%
  pivot_wider(names_from = "time",  #Preparing for use of rowMeans
              values_from = "libdem")

merged_time1 <- merged_time1 %>%
  mutate(`dem_1981_change` = `1981`-`1980`,
         `dem_1982_change` = `1982`-`1981`,
         `dem_1983_change` = `1983`-`1982`,
         `dem_1984_change` = `1984`-`1983`,
         `dem_1985_change` = `1985`-`1983`,
         `dem_1986_change` = `1986`-`1985`,
         `dem_1987_change` = `1987`-`1986`,
         `dem_1988_change` = `1988`-`1987`,
         `dem_1989_change` = `1989`-`1988`,
         `dem_1990_change` = `1990`-`1989`,
         `dem_1991_change` = `1991`-`1990`,
         `dem_1992_change` = `1992`-`1991`,
         `dem_1993_change` = `1993`-`1992`,
         `dem_1994_change` = `1994`-`1993`,
         `dem_1995_change` = `1995`-`1994`,
         `dem_1996_change` = `1996`-`1995`,
         `dem_1997_change` = `1997`-`1996`,
         `dem_1998_change` = `1998`-`1997`,
         `dem_1999_change` = `1999`-`1998`,
         `dem_2000_change` = `2000`-`1999`,
         `dem_2001_change` = `2001`-`2000`,
         `dem_2002_change` = `2002`-`2001`,
         `dem_2003_change` = `2003`-`2002`,
         `dem_2004_change` = `2004`-`2003`,
         `dem_2005_change` = `2005`-`2004`,
```

```r
         `dem_2006_change` = `2006`-`2005`,
         `dem_2007_change` = `2007`-`2006`,
         `dem_2008_change` = `2008`-`2007`,
         `dem_2009_change` = `2009`-`2008`,
         `dem_2010_change` = `2010`-`2009`,
         `dem_2011_change` = `2011`-`2010`,
         `dem_2012_change` = `2012`-`2011`,
         `dem_2013_change` = `2013`-`2012`,
         `dem_2014_change` = `2014`-`2013`,
         `dem_2015_change` = `2015`-`2014`,
         `dem_2016_change` = `2016`-`2015`,
         `dem_2017_change` = `2017`-`2016`,
         `dem_2018_change` = `2018`-`2017`,
         `dem_2019_change` = `2019`-`2018`,
         `dem_2020_change` = `2020`-`2019`)

merged_time1  <- merged_time1  %>%
  select(1, 43:82)

# Gini Index before tax

merged_time2 <- merged_data1 %>%
  filter(!(str_detect(time, "past|present"))) %>%
  mutate(time = as.numeric(time)) %>%
  filter(time >= 1980) %>%
  select(country, time, gini_before_tax) %>%
  drop_na() %>%
  pivot_wider(names_from = "time",  #Preparing for use of rowMeans
              values_from = "gini_before_tax")

merged_time2 <- merged_time2 %>%
  mutate(`gini_1981_change` = `1981`-`1980`,
         `gini_1982_change` = `1982`-`1981`,
         `gini_1983_change` = `1983`-`1982`,
         `gini_1984_change` = `1984`-`1983`,
         `gini_1985_change` = `1985`-`1984`,
         `gini_1986_change` = `1986`-`1985`,
         `gini_1987_change` = `1987`-`1986`,
         `gini_1988_change` = `1988`-`1987`,
         `gini_1989_change` = `1989`-`1988`,
         `gini_1990_change` = `1990`-`1989`,
         `gini_1991_change` = `1991`-`1990`,
         `gini_1992_change` = `1992`-`1991`,
         `gini_1993_change` = `1993`-`1992`,
         `gini_1994_change` = `1994`-`1993`,
         `gini_1995_change` = `1995`-`1994`,
         `gini_1996_change` = `1996`-`1995`,
         `gini_1997_change` = `1997`-`1996`,
         `gini_1998_change` = `1998`-`1997`,
         `gini_1999_change` = `1999`-`1998`,
```

```r
         `gini_2000_change` = `2000`-`1999`,
         `gini_2001_change` = `2001`-`2000`,
         `gini_2002_change` = `2002`-`2001`,
         `gini_2003_change` = `2003`-`2002`,
         `gini_2004_change` = `2004`-`2003`,
         `gini_2005_change` = `2005`-`2004`,
         `gini_2006_change` = `2006`-`2005`,
         `gini_2007_change` = `2007`-`2006`,
         `gini_2008_change` = `2008`-`2007`,
         `gini_2009_change` = `2009`-`2008`,
         `gini_2010_change` = `2010`-`2009`,
         `gini_2011_change` = `2011`-`2010`,
         `gini_2012_change` = `2012`-`2011`,
         `gini_2013_change` = `2013`-`2012`,
         `gini_2014_change` = `2014`-`2013`,
         `gini_2015_change` = `2015`-`2014`,
         `gini_2016_change` = `2016`-`2015`,
         `gini_2017_change` = `2017`-`2016`,
         `gini_2018_change` = `2018`-`2017`,
         `gini_2019_change` = `2019`-`2018`,
         `gini_2020_change` = `2020`-`2019`)

merged_time2  <- merged_time2  %>%
  select(1, 43:82)

# Merging data

merged_time <- merged_time1 %>%
  full_join(merged_time2,
            by = "country") %>%
  arrange(desc(country)) %>%    #Sorting alphabetically
  drop_na() %>%  #Dropping missing values
  pivot_longer(c(-country),
               names_to = "names",
               values_to = "values") %>%
  mutate(year = extract_numeric(names),
         variable = map(str_split(names, "_"),
                        function(x){x[1]})
         ) %>%
  select(-names) %>%
  pivot_wider(names_from = "variable",
              values_from = "values") %>%
  mutate(change = NA) %>%
  mutate(change =
           case_when(gini <= 0 & dem >= 0 ~ 1, #More democratic and more equal
                     gini >= 0 & dem >= 0 ~ 2, #More democratic and less equal
                     gini <= 0 & dem <= 0 ~ 3, #Less democratic and more equal
                     gini >= 0 & dem <= 0 ~ 4)) #Less democratic and less equal
```

```
## extract_numeric() is deprecated: please use readr::parse_number() instead
```

Visualization

Line graph

```
# Data frame for the plot

time_data <- merged_time %>%
  select(change, year) %>%
  group_by(change, year) %>%
  tally() %>%
  pivot_wider(names_from = "change",
              values_from = "n") %>%
  rename(more_dem_more_equal = 2,
         more_dem_less_equal = 3,
         less_dem_more_equal = 4,
         less_dem_less_equal = 5)

# Colors

colors <- c("More democratic and more equal" = "dodgerblue4",
            "More democratic and less equal"  = "dodgerblue1",
            "Less democratic and more equal"  = "firebrick1",
            "Less democratic and less equal"  = "firebrick4")

# Time series plot

plot <- ggplot(time_data,
            aes(x = year))+
  geom_line(aes(y = more_dem_more_equal,
                color = "More democratic and more equal"),
            size=1,
            alpha=0.5,
            linetype="twodash") +
  geom_line(aes(y = more_dem_less_equal,
                color = "More democratic and less equal"),
            size=1,
            alpha=0.5,
            linetype="twodash")+
  geom_line(aes(y = less_dem_more_equal,
                color = "Less democratic and more equal"),
            size=1,
            alpha=0.5,
            linetype="twodash") +
  geom_line(aes(y = less_dem_less_equal,
                color = "Less democratic and less equal"),
            size=1,
            alpha=0.5,
            linetype="twodash")+
  geom_smooth(aes(y=more_dem_more_equal),
            geom="line",
```

```
          method="loess",
          color="dodgerblue4",
          fill="dodgerblue4",
          size=1.5,
          level=0.9,
          se = FALSE)+
geom_smooth(aes(y=more_dem_less_equal),
          geom="line",
          method="loess",
          color= "dodgerblue1",
          fill= "dodgerblue1",
          size=1.5,
          level=0.9,
          se = FALSE)+
geom_smooth(aes(y=less_dem_more_equal),
          geom="line",
          method="loess",
          color="firebrick1",
          fill="firebrick1",
          size=1.5,
          alpha=0.1,
          level=0.9,
          se = FALSE)+
geom_smooth(aes(y=less_dem_less_equal),
          geom="line",
          method="loess",
          color= "firebrick4",
          fill= "firebrick4",
          size=1.5,
          level=0.9,
          se = FALSE)+
scale_x_continuous(limits = c(1980, 2020),
                   breaks = seq(1980, 2020, by = 5),
                   minor_breaks = seq(1980, 2020, by = 1))+
scale_y_continuous(limits = c(0, 85),
                   breaks = seq(0, 85, by = 10))+
scale_color_manual(values = colors,
                   labs(fill = "Country change from last year"))+
labs(y = "\nNumber of countries\n",
     x = element_blank(),
     title = "A third wave of autocratization",
     subtitle = "Year-over-year change in democracy and economic equality in 140 countries",
     caption = "\nSource: Liberal Democracy Index from 'Varieties of Democracy Institute', Gini
theme_bw()+
theme(strip.text = element_text(size = 15),
      axis.title = element_text(size = 20),
      axis.text.x = element_text(size = 15),
      axis.text.y = element_text(size = 15),
      legend.title = element_blank(),
      legend.text = element_text(size = 15),
```

```
        legend.key.height = unit(1, 'cm'),
        plot.title = element_text(size = 25,
                                  margin=margin(20,0,0,0)),
        plot.subtitle = element_text(size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5)
  )
```

```
## Warning: Ignoring unknown parameters: geom
## Ignoring unknown parameters: geom
## Ignoring unknown parameters: geom
## Ignoring unknown parameters: geom
```

```
# Showing and saving plot

plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

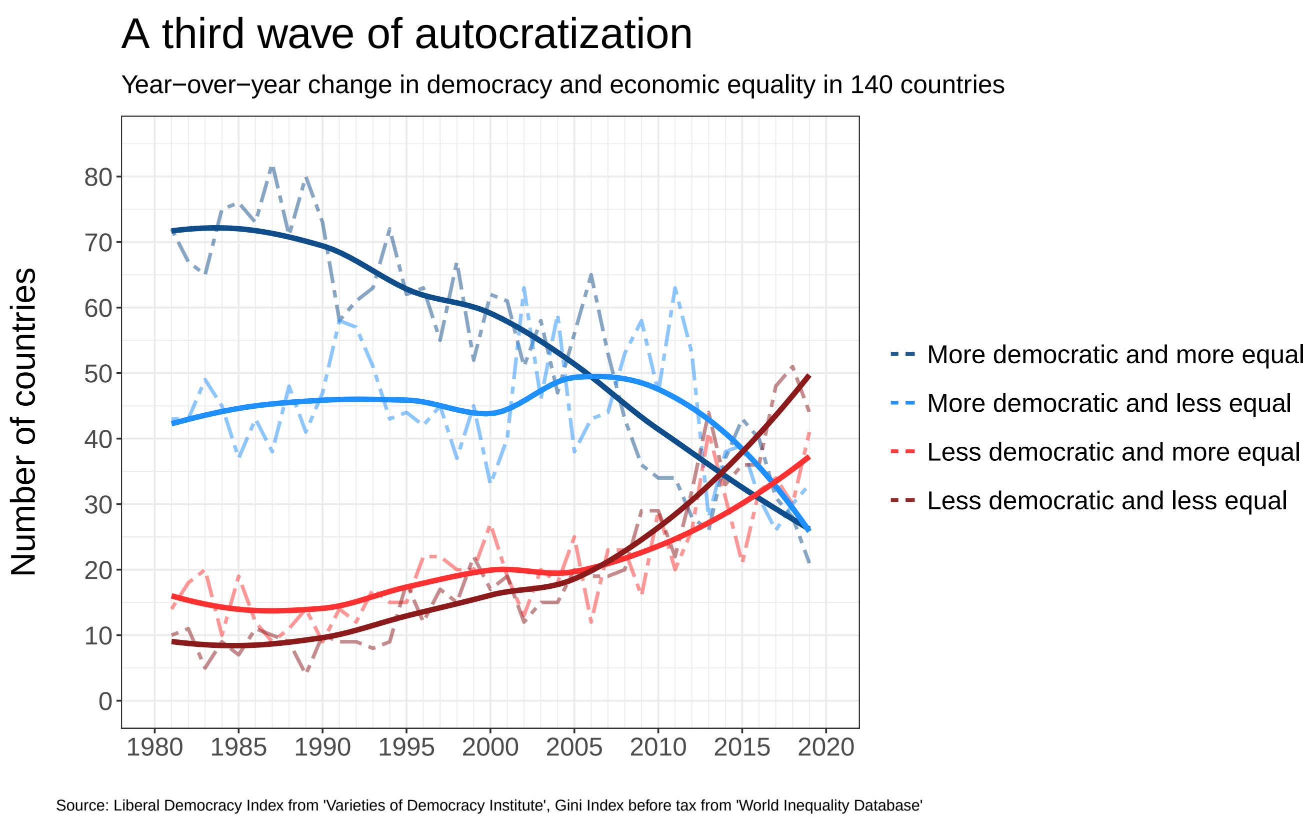
```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
## Removed 2 row(s) containing missing values (geom_path).
## Removed 2 row(s) containing missing values (geom_path).
## Removed 2 row(s) containing missing values (geom_path).
```




```
ggsave("special_plot.png", width = 11, height = 8)
```

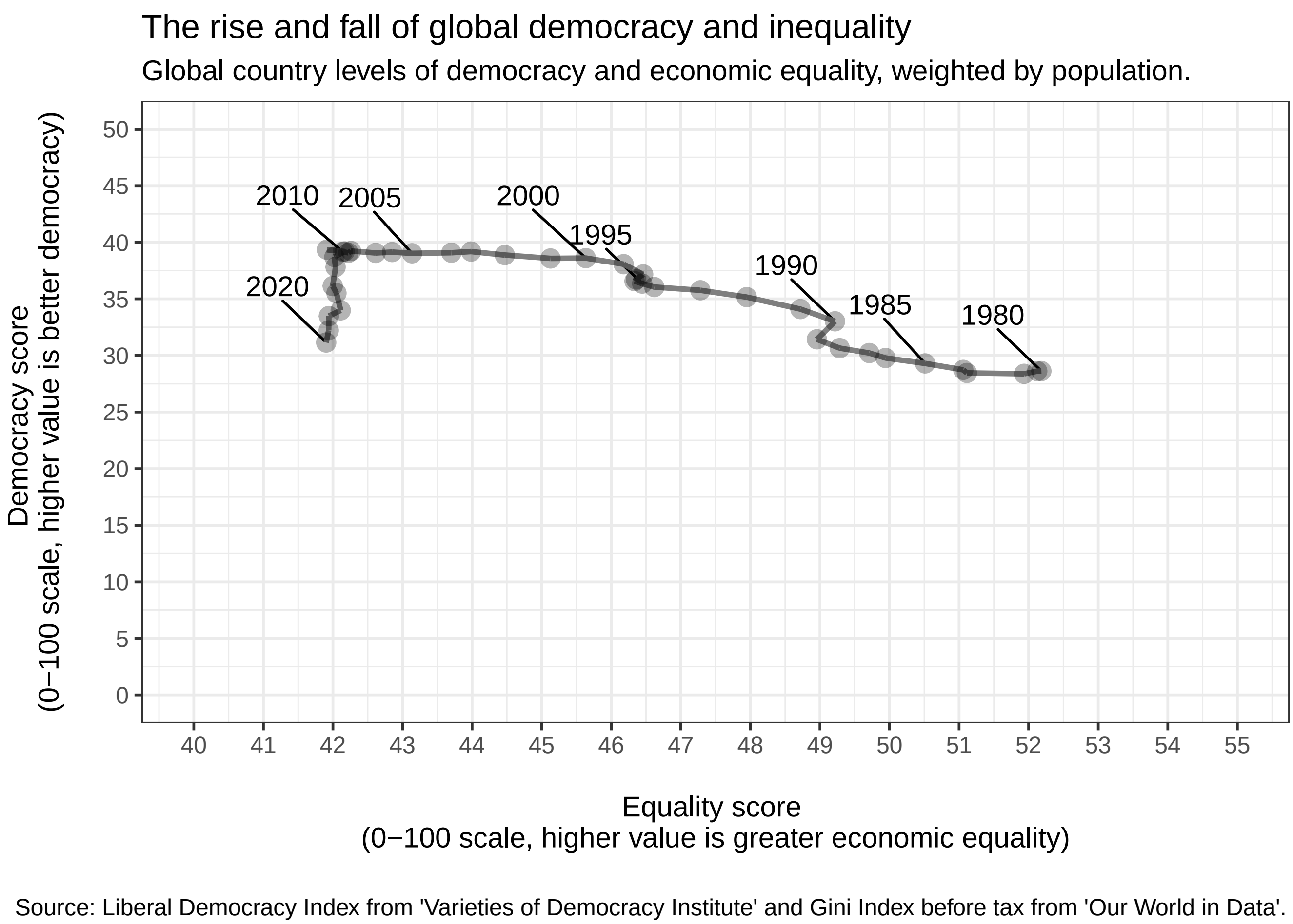
```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 2 rows containing non-finite values (stat_smooth).
## Removed 2 row(s) containing missing values (geom_path).
## Removed 2 row(s) containing missing values (geom_path).
## Removed 2 row(s) containing missing values (geom_path).
## Removed 2 row(s) containing missing values (geom_path).
```

The influence of the noeliberal era and globalization of world politics can be seen from 2000 to 2010, when we observe a large number of countries becoming both less equal in the democratic world. The impact of the financial crisis can be spotted, from the sudden decrease in the number of more equal and democratic countries after 2008. A general decline of democracy is evident after turn of the century. Recently, this "third wave" of autocratization has been primarily driven by countries becoming more authoritarian and less equal at the same time. It is true that more countries are becoming less democratic and more equal, but this is still a very recent trend. Hypothesis 3 is correct, but not significantly relevant.

Connected scatter plot

```r
# Time series data

data_world <- merged_data1 %>%
  filter(!(str_detect(time, "past|present"))) %>%
  mutate(time = as.numeric(time)) %>%
  filter(time > 1979) %>%
  select(country, time, gini_before_tax, libdem, population) %>%
  drop_na() %>%
  group_by(time) %>%
  summarise(gini = weighted.mean(gini_before_tax, population),
            libdem = weighted.mean(libdem, population))

# Time series plot

plot <- ggplot(data_world,
               aes(x = gini, y = libdem, label = time))+
  geom_text_repel(min.segment.length = 0,
                  nudge_y = 5,
                  nudge_x = -0.7,
                  data = subset(data_world,
                                time %in% c(1980, 1985, 1990, 1995,
                                            2000, 2005, 2010, 2020))
                  ) +
  geom_segment(color = "grey50",
               size = 1,
               aes(xend=c(tail(gini, n=-1), NA),
                   yend=c(tail(libdem, n=-1), NA))
               ) +
  geom_point(size = 3, alpha = 0.3) +
  scale_y_continuous(limits = c(0, 50), breaks = seq(0, 50, by = 5)) +
  scale_x_continuous(limits = c(40, 55), breaks = seq(40, 55, by = 1)) +
  labs(x = "\nEquality score \n(0-100 scale, higher value is greater economic equality)\n",
       y = "Democracy score \n(0-100 scale, higher value is better democracy)\n",
       title = "The rise and fall of global democracy and inequality",
       subtitle = "Global country levels of democracy and economic equality, weighted by populatio
       caption = "Source: Liberal Democracy Index from 'Varieties of Democracy Institute' and Gini
  theme_bw()

plot
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```




```r
ggsave("world_plot_weighted.png", width = 7, height = 5)
```

```
## Warning: Removed 1 rows containing missing values (geom_segment).
```

After thirty years of democratic progress and falling economic inequality around the world, the trends have shifted. On the plus side, a thirty-year decline in economic equality within countries, have finally been stopped. The global levels of economic equality have been remarkably stable since the financial crisis of 2008, albeit on a relatively low level of 42. On the flip side, the global levels of democracy have dramatically fallen, from 39 in 2008 to 31 in 2020. All the democratic gains since the fall of the Berlin Wall in 1989, seem to have been erased.

8 Conclusion

The recent wave of autocratization in the world has prompted increased attention to the causes behind democracy. Economic equality is imperative for maintaining and fostering a healthy democracy. Evidence suggests that equal countries are, on average, more democratic than unequal ones. This partially approves of hypothesis 2. After controlling for the effects of GDP per capita, planned economies, and dependence on natural resources, equality appears to be correlated with democracy. This relationship is true for every year since at least 1990. The relatively simple economic model presented in this paper, is able to explain a moderately high amount of the variation in democratic quality around the world. Nevertheless, when economic equality is measured after tax, it does not seem to have sizeable impact on democracy. Hypothesis 1 is neither sufficiently approved nor disapproved. Whether or not a country is becoming increasingly equal over time, cannot predict if the country is undergoing democratization. This disapproves of hypothesis 3. The third wave of autocratization is for the most part driven by countries becoming both decreasingly equal and democratic, but a significant number of countries are in fact becoming more equal and less democratic. This partially approves of hypothesis 4. To sum things up, there is much to indicate that economic equality has a positive impact on democracy, but that democratization over time cannot be solely explained by economic factors.

9 Appendix

In this final section, I provide some interesting results from the analysis, which fall outside the scope of the hypotheses. This analysis has shown that countries differ drastically from each other. Each country is different from one another, but some countries are more different than others. I try to measure exactly which country is the most similar to other countries in the world. I have done this by ranking all countries, according to how they score in certain indicators measuring the way of life. The indicators I chosen are per capita income, social inequality, and democratic quality.

I have taken the avarage score of these indicators across all the countries and calculated each country's absolute distance to these avarage values (regardless if it is positive or negative). I have weighted all the indicators equally, by dividing the country distances with the standard deviation of each indicator. This is what statisticians call z-scores, a fancy name for standardized variables. I then added up the absolute values each country's z-score on every variable, to get a final rank of all countries in the world, according to how similar they are to the world mean.

What is the most avarage country in the world?

```
# Findind the avarage values of three indicators measuring the way of life

avarage_world <- merged_data1 %>%
  filter(time == "2019") %>%
  mutate(abs_libdem = abs(libdem - mean(libdem, na.rm = TRUE)),
         abs_gini_after_tax = abs(gini_after_tax - mean(gini_after_tax, na.rm = TRUE)),
         abs_gdpcapita = abs(gdpcapita - mean(gdpcapita, na.rm = TRUE)),
         sd_libdem = sd(libdem, na.rm = TRUE),
         sd_gini_after_tax = sd(gini_after_tax, na.rm = TRUE),
         sd_gdpcapita = sd(gdpcapita, na.rm = TRUE),
         z_score = (abs_libdem/sd_libdem)+
           (abs_gini_after_tax/sd_gini_after_tax)+
           (abs_gdpcapita/sd_gdpcapita),
         score_width = cut_width(z_score, width = 0.25, boundary = 0)
         )

mean(avarage_world$gini_after_tax, na.rm = TRUE)   #62
```

```
## [1] 61.965
```

```r
mean(avarage_world$libdem, na.rm = TRUE)    #41
```

```
## [1] 41.05607
```

```r
mean(avarage_world$gdpcapita, na.rm = TRUE)    #20 533
```

```
## [1] 20533.83
```

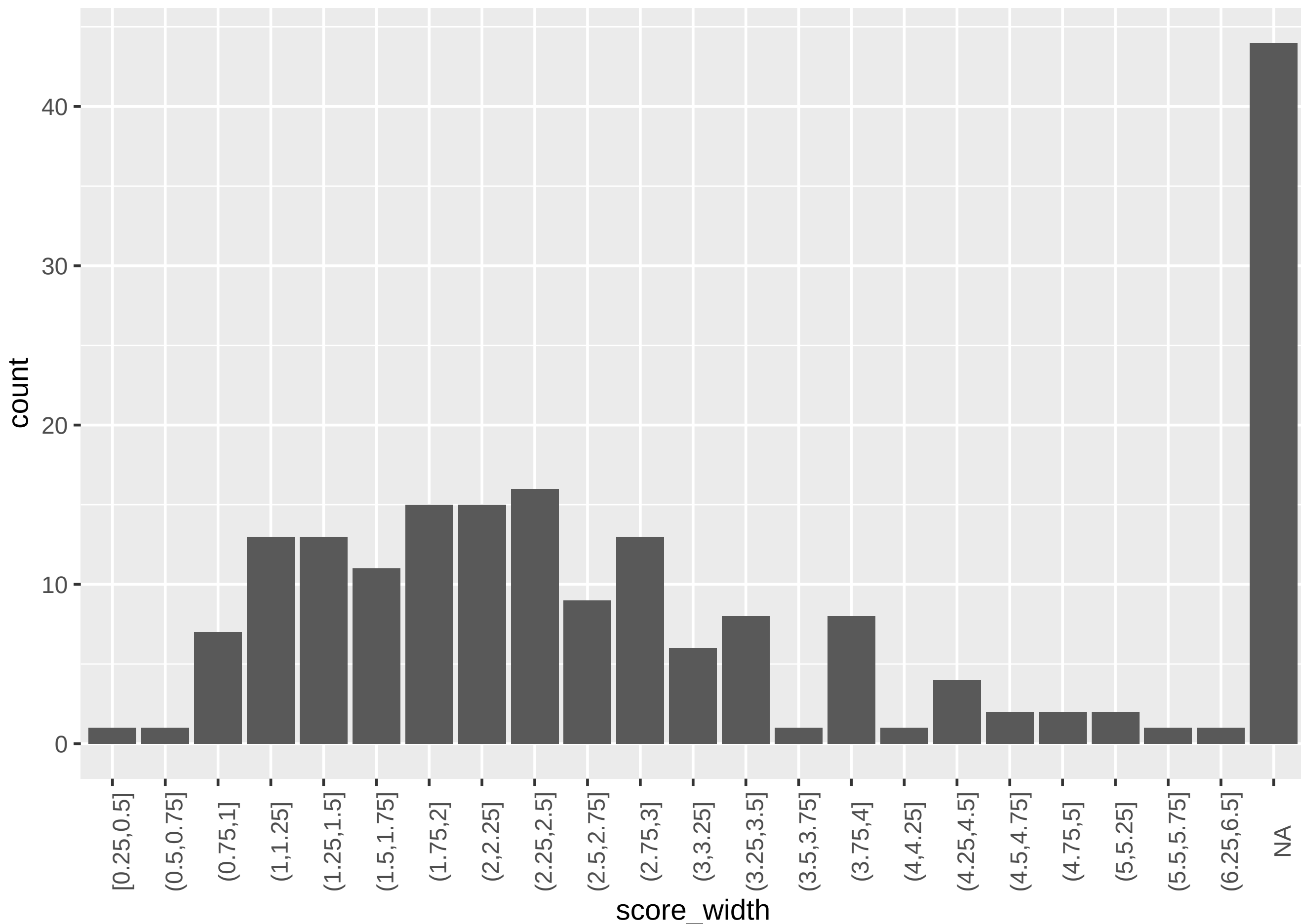
```r
# Finding the countries closest representing the world

avarage_world %>%
  filter(population > 10^6) %>%     #Excluding small states
  select(country, gini_after_tax, libdem, gdpcapita, z_score) %>%
  arrange(z_score)
```

```
## # A tibble: 154 x 5
##    country         gini_after_tax libdem gdpcapita z_score
##    <chr>                    <dbl>  <dbl>     <dbl>   <dbl>
##  1 Sri Lanka                 60.7   44.2    13070.   0.645
##  2 El Salvador               61.2   43.8     8776.   0.764
##  3 Indonesia                 60.8   47.6    11812.   0.819
##  4 North Macedonia           67.0   42.2    16773.   0.871
##  5 Georgia                   64.1   51.2    14989.   0.920
##  6 Albania                   66.8   40.7    13657.   0.964
##  7 Bulgaria                  58.7   52.1    23266.   0.968
##  8 Philippines               62.2   30       8915.   1.01
##  9 Gabon                     62.0   21      14946.   1.04
## 10 Malaysia                  58.9   33.8    28421.   1.04
## # ... with 144 more rows
```

```r
avarage_world %>%
  filter(population > 10^6) %>% #Excluding  small states
  select(country, gini_after_tax, libdem, gdpcapita, z_score) %>%
  arrange(-z_score)
```

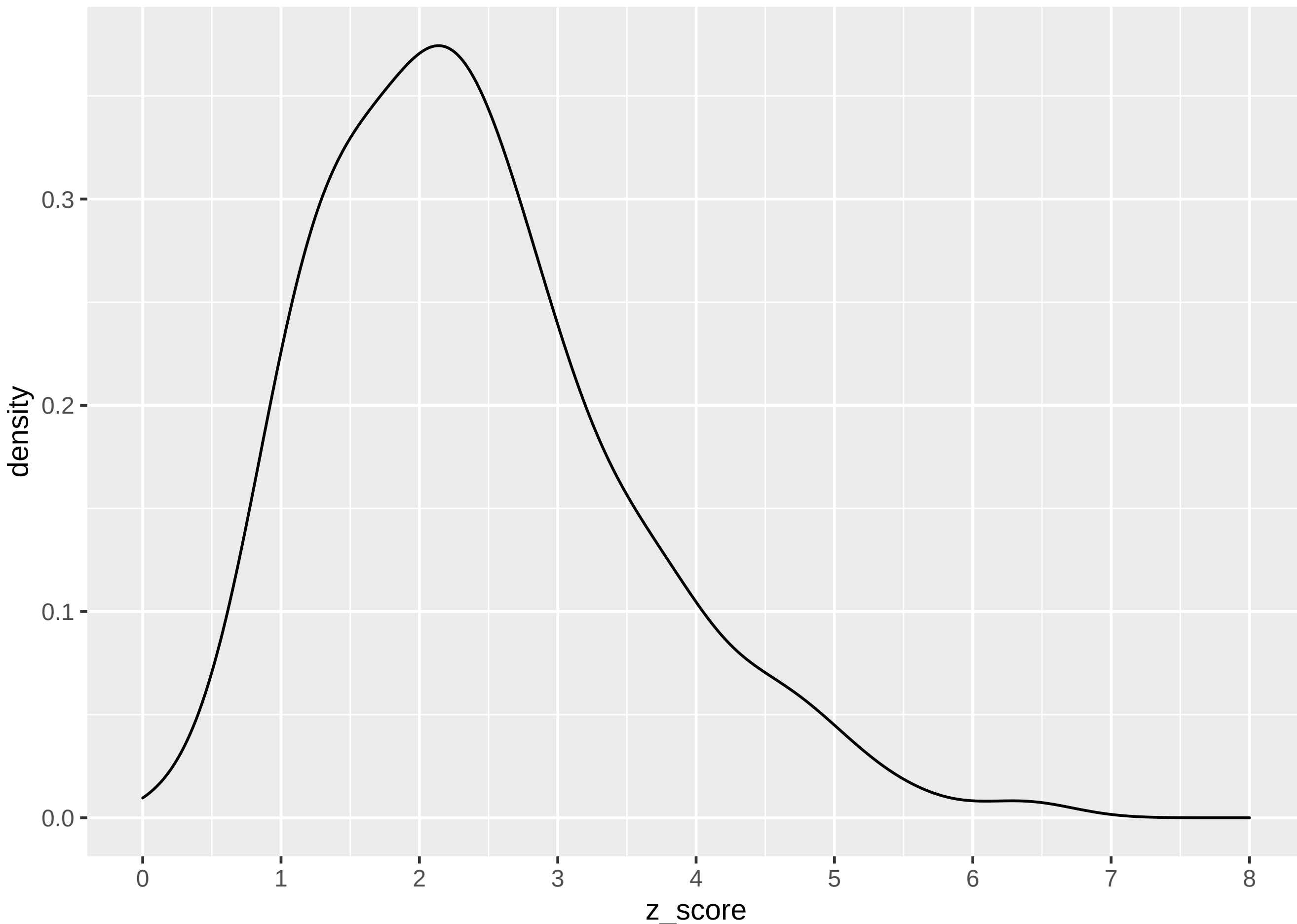
```
## # A tibble: 154 x 5
##    country              gini_after_tax libdem gdpcapita z_score
##    <chr>                         <dbl>  <dbl>     <dbl>   <dbl>
##  1 Ireland                        68.6   81      86650.    5.55
##  2 Norway                         72.4   85.5    64453.    5.14
##  3 United Arab Emirates           74.0    9      68264.    5.05
##  4 Denmark                        71.8   88.3    57162.    4.82
##  5 Switzerland                    66.9   85.9    70944.    4.76
##  6 Netherlands                    71.9   82.3    56784.    4.58
##  7 Belgium                        72.8   82.8    51944.    4.48
##  8 Sweden                         70.0   87.9    52851.    4.37
##  9 Finland                        72.7   83.3    48642.    4.33
## 10 South Africa                   37.0   60.9    13710.    4.27
## # ... with 144 more rows
```

```r
# Plotting the distrubtion of the combined z-scores

ggplot(avarage_world, aes(score_width)) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90))   #Rotating x-axis text labels
```




```r
ggplot(avarage_world, aes(z_score))+
  geom_density(kernel = "gaussian")+
  scale_x_continuous(limits = c(0, 8),
                     breaks = seq(0, 8, by = 1))
```

```
## Warning: Removed 44 rows containing non-finite values (stat_density).
```

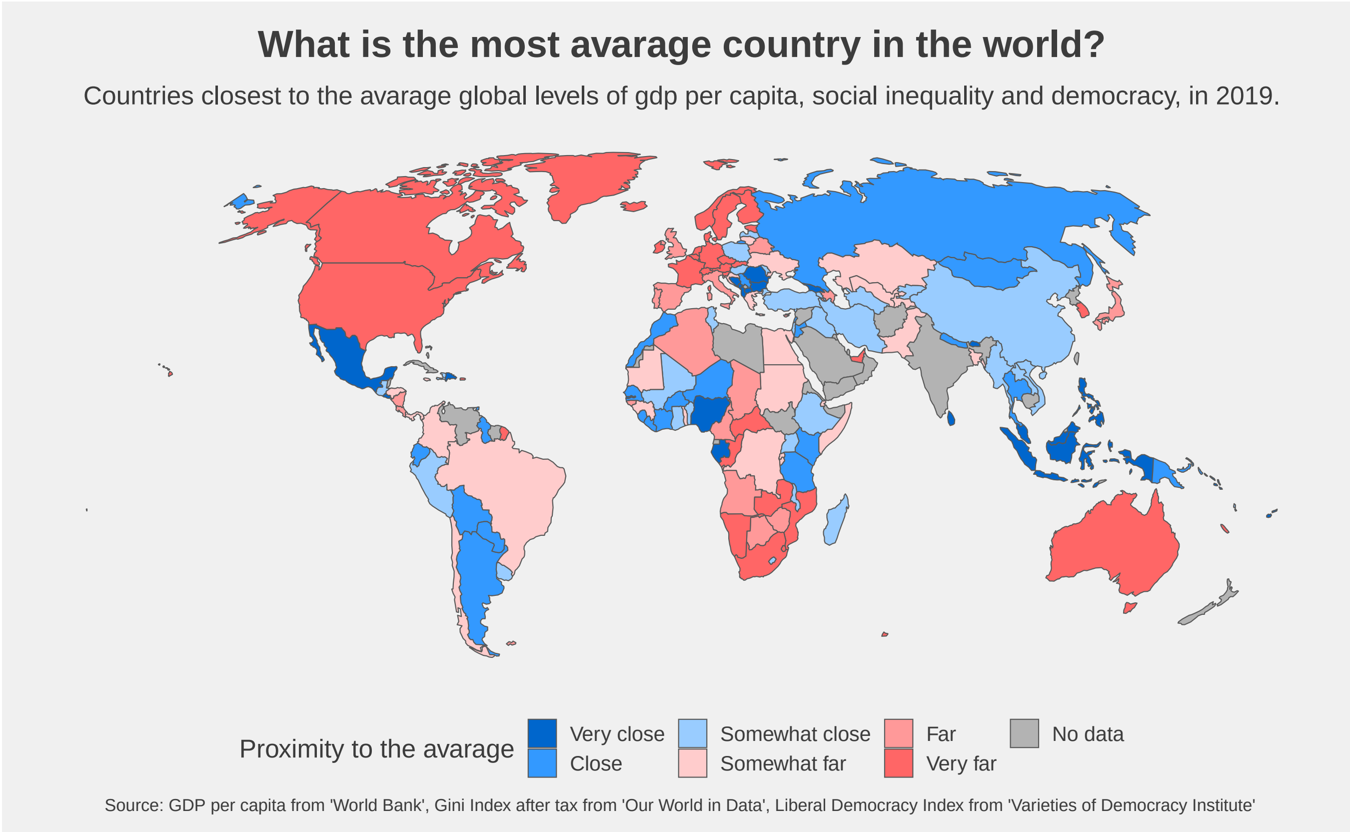


Sri Lanka, El Salvador and Indonesia are the most avarage countries in thw world. They come the closest to the avarage global levels of gdp per capita, social inequality, and democratic quality. Ireland, Norway and the United Arab Emirates are the most unusual countries in the world, but for totally different reasons. Ireland and Norway are richer and more democratic than the rest of the world. The United Arab Emirates on the other hand, have a exceptionally low level of democracy.

```
# Map data

map_data3 <- raw_map_data %>%
  select(sovereignt, geometry) %>%
  filter(sovereignt != "Antarctica") %>%
  rename(country = sovereignt) %>%
  mutate(id = countrycode(country,
                          origin = "country.name",
                          destination = "iso3c")) %>%
  left_join(avarage_world, by = "id") %>%
  mutate(combination = cut(z_score,
                           breaks = c(-Inf, 1.25, 1.75, 2.25, 2.75, 3.25, Inf),
                           labels = c("Very close",
                                      "Close",
                                      "Somewhat close",
                                      "Somewhat far",
                                      "Far",
                                      "Very far")))
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```
# Map colors

map4_colors <- c("Very close" = "#0066CC",
                 "Close"= "#3399FF",
                 "Somewhat close" = "#99CCFF",
                 "Somewhat far" = "#FFCCCC",
                 "Far" = "#FF9999",
                 "Very far" = "#FF6666",
                 "No data" = "gray70")

# Map making

map4 <- ggplot(data = map_data3) +
  geom_sf(aes(fill = combination), lwd = 0.3) +   #Adjusting border size
  scale_fill_manual(values = map4_colors,   #Defined colors
                    na.value = "gray70",   #Color missing countries grey
                    labs(fill = "Proximity to the avarage"))+   #Legend title
  coord_sf(crs ="+proj=robin +lon_0=0 +x_0=0   #A better map projection
           +y_0=0 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")+
  labs(title = "What is the most avarage country in the world?",
       subtitle = "Countries closest to the avarage global levels of gdp per capita, social inequa
       caption = "Source: GDP per capita from 'World Bank', Gini Index after tax from 'Our World i
  theme_fivethirtyeight()+   #Pleasent theme
  theme(panel.grid.major = element_blank(),   #Removing longditude and latitude
        axis.text = element_blank(),   #No unnecessary titles
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(2,0,2,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5),
        legend.title = element_text(size = 15),
        legend.text = element_text(size = 12))+
  guides(fill = guide_legend(ncol = 4))

# Showing and saving map

map4
```




```
ggsave("map4.png", width = 11, height = 7)
```

Southeastern Asia seems to be the most average world region. Countries like Indonesia, Sri Lanka and the Philippines are all middle-income countries which hold regular elections. They don't score the best in terms of equality or democratic norms, which closes the gap to the average. Southern Africa fare rather poorly, in part because the former is a region with much lower levels of equality and income than global average. But there are exceptions, Gabon and Mauritius for example, come close to the global income average with their huge oil and tourist industry, respectively.

What is the most avarage country in Europe?

I conduct the exact same analysis as above but restrict the sample to countries in Europe.

```
# Findind the avarage values of three indicators measuring the way of life

avarage_europe <- merged_data1 %>%
  filter(time == "2019", Continent == "Europe") %>%
  mutate(abs_libdem =
           abs(libdem - mean(libdem, na.rm = TRUE)),
         abs_gini_after_tax =
           abs(gini_after_tax - mean(gini_after_tax, na.rm = TRUE)),
         abs_gdpcapita =
           abs(gdpcapita - mean(gdpcapita, na.rm = TRUE)),
         sd_libdem = sd(libdem, na.rm = TRUE),
         sd_gini_after_tax = sd(gini_after_tax, na.rm = TRUE),
```

```
         sd_gdpcapita = sd(gdpcapita, na.rm = TRUE),
         z_score =
           (abs_libdem/sd_libdem)+
           (abs_gini_after_tax/sd_gini_after_tax)+
           (abs_gdpcapita/sd_gdpcapita),
         score_width = cut_width(z_score, width = 0.25, boundary = 0)
         )

mean(avarage_europe$gini_after_tax, na.rm = TRUE)   #68.57
```

```
## [1] 68.57458
```

```
mean(avarage_europe$libdem, na.rm = TRUE)   #61.83
```

```
## [1] 61.83182
```

```
mean(avarage_europe$gdpcapita, na.rm = TRUE)   #38 947
```

```
## [1] 38947.48
```

```
# Finding the countries closest representing Europe

avarage_europe %>%   #The countries closest to the avarage
  select(country, gini_after_tax, libdem, gdpcapita, z_score) %>%
  arrange(z_score)
```

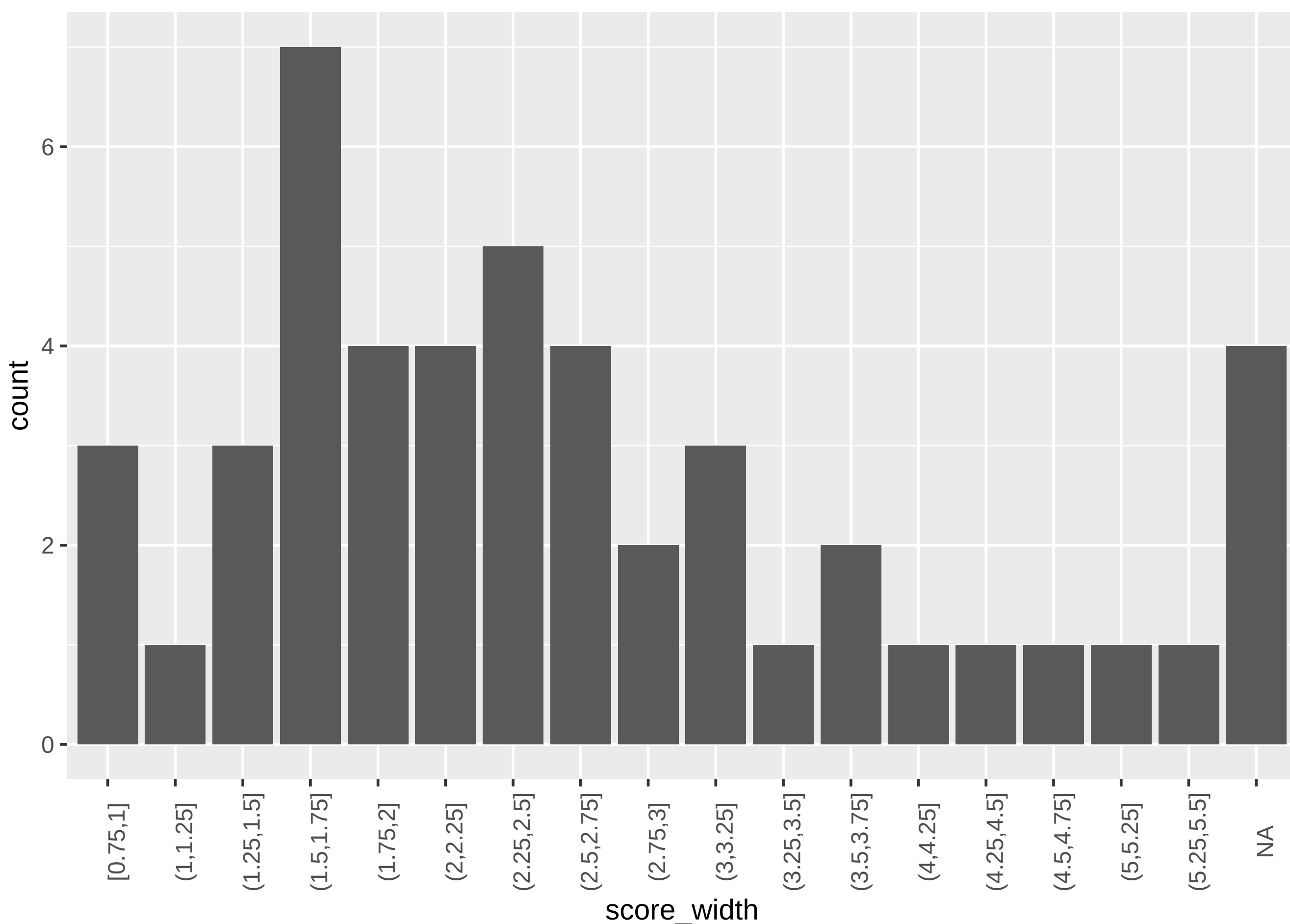
```
## # A tibble: 48 x 5
##    country  gini_after_tax libdem gdpcapita z_score
##    <chr>             <dbl>  <dbl>     <dbl>   <dbl>
##  1 Cyprus             67.3   72.8    40227.   0.814
##  2 Croatia            70.3   61.9    29336.   0.849
##  3 Poland             69.8   51.9    33185.   0.955
##  4 Malta              71.3   57.6    43951.   1.05
##  5 Estonia            69.7   83.9    36400.   1.28
##  6 France             67.6   80      46018.   1.30
##  7 Greece             67.1   76      29698.   1.36
##  8 Austria            69.2   76.5    55834.   1.54
##  9 Portugal           66.5   84.5    34946.   1.60
## 10 Armenia            70.1   63.2    13654.   1.61
## # ... with 38 more rows
```

```
avarage_europe %>%   #The countries furthest from the avarage
  select(country, gini_after_tax, libdem, gdpcapita, z_score) %>%
  arrange(-z_score)
```

```
## # A tibble: 48 x 5
##    country    gini_after_tax libdem gdpcapita z_score
```

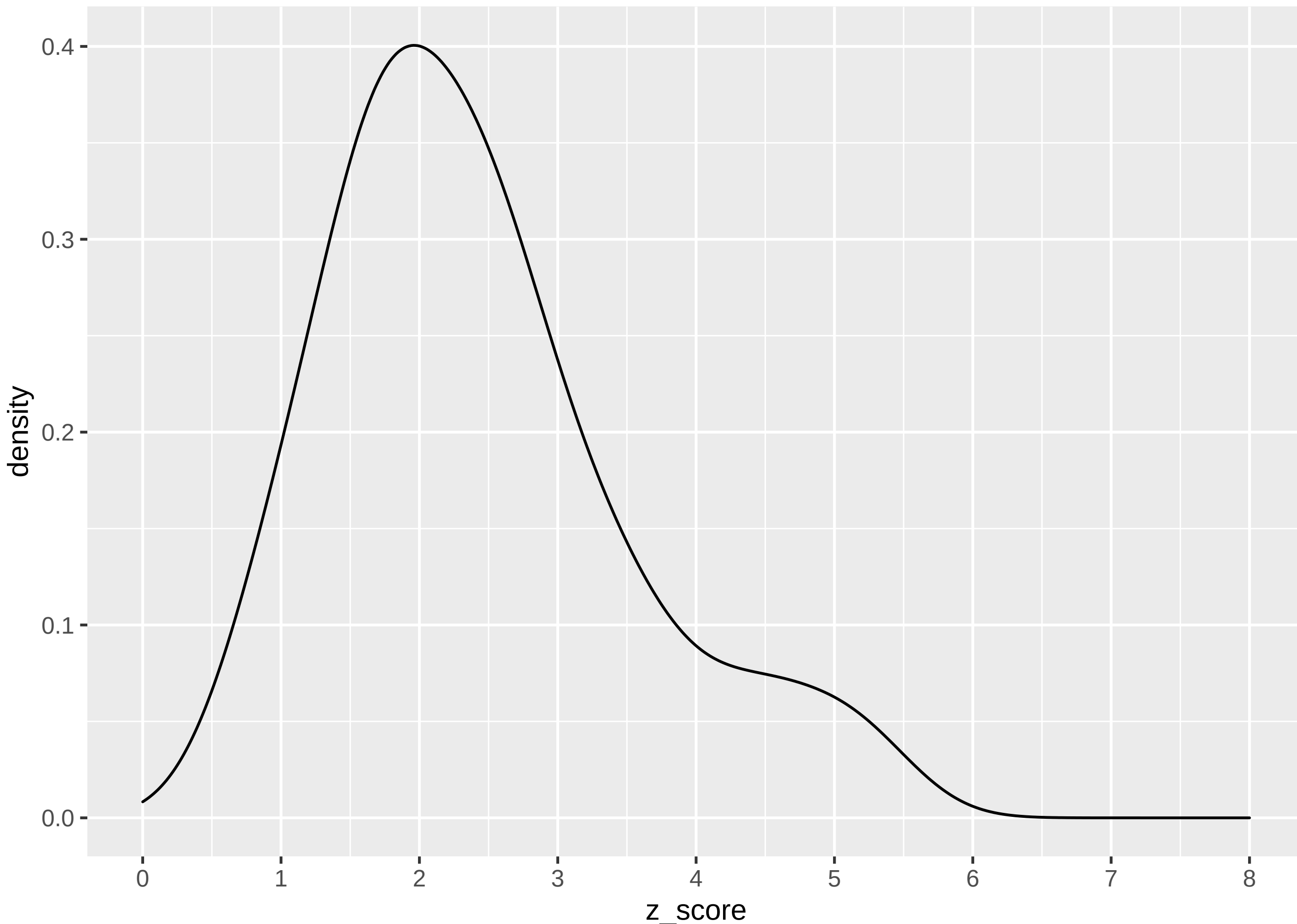
```
##    <chr>                <dbl>  <dbl>     <dbl>   <dbl>
##  1 Luxembourg            64.6   78.6   116518.    5.28
##  2 Turkey                58.1   10.8    28197.    5.02
##  3 Azerbaijan            73.4    6.1    14442.    4.58
##  4 Belarus               74.7   11.4    19279.    4.42
##  5 Russia                62.5   10.9    27211.    4.04
##  6 Ukraine               73.4   29.2    12809.    3.69
##  7 Montenegro            61.5   35.1    21534.    3.55
##  8 Bulgaria              58.7   52.1    23266.    3.41
##  9 Moldova               74.3   45.6    13027.    3.22
## 10 Norway                72.4   85.5    64453.    3.07
## # ... with 38 more rows
```

```r
# Plotting the distrubution of the combined z-scores

ggplot(avarage_europe, aes(score_width))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90))   #Rotating x-axis text labels
```




```r
ggplot(avarage_europe, aes(z_score))+
  geom_density(kernel = "gaussian")+
  scale_x_continuous(limits = c(0, 8),
                     breaks = seq(0, 8, by = 1))
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```



Cyprus, Croatia and Poland are the most normal European countries. They come the closest to the avarage European levels of per capita income, social inequality, and democratic quality. Azerbaijan, Turkey and Luxembourg are the most unusual European countries, but for totally different reasons. Azerbaijan and Turkey are much poorer and less democratic than the rest of Europe. Luxembourg on the other hand, have exceptionally high and unequally distributed levels of income.

```
# Europe map

world <- ne_countries(scale = "medium", type = 'sovereignty', returnclass = "sf")

raw_europe_data <- world[which(world$region_wb == "Europe & Central Asia"),]

map_data4 <- raw_europe_data %>%
  select(sovereignt, geometry) %>%
  filter(sovereignt != "Kazakhstan") %>% # Kazakhstan not part of Europe
  rename(country = sovereignt) %>%
  mutate(id = countrycode(country,
                          origin = "country.name",
                          destination = "iso3c")) %>%
  left_join(avarage_europe, by = "id") %>%
  mutate(combination = cut(z_score, #Ordinal variable by 0.5 standard deviations
```

```r
                            breaks = c(-Inf, 1.25, 1.75, 2.25, 2.75, 3.25, Inf),
                            labels = c("Very close",
                                       "Close",
                                       "Somewhat close",
                                       "Somewhat far",
                                       "Far",
                                       "Very far")))
```

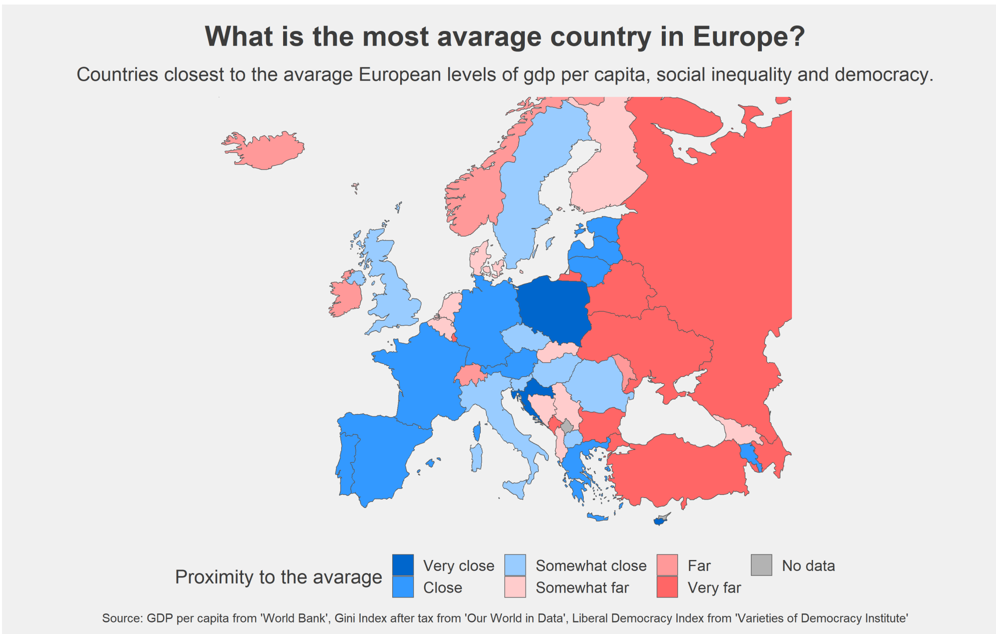
```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : So
```

```r
map_data4[11, 26] <- "NA" #Making Northern Cyprus NA
```

```
## Warning in `[<-.factor`(`*tmp*`, iseq, value = "NA"): invalid factor level, NA
## generated
```

```r
# Map colors

map5_colors <- c("Very close" = "#0066CC",
                 "Close"= "#3399FF",
                 "Somewhat close" = "#99CCFF",
                 "Somewhat far" = "#FFCCCC",
                 "Far" = "#FF9999",
                 "Very far" = "#FF6666",
                 "No data" = "gray70")

# Map making

map5 <- ggplot(data = map_data4) +
  geom_sf(aes(fill = combination),   #Coloring by ordinal values
          lwd = 0.3) +   #Reducing border size
  coord_sf(xlim = c(-24.8, 50.4),   #Zooming in from default; ylim c(35.8, 69.3)
           ylim = c(34.4, 69.3), expand = FALSE)+
  scale_fill_manual(values = map5_colors,   #Blue/red colors
                    na.value = "gray70",   #Color missing countries grey
                    labs(fill = "Proximity to the avarage")) +   #Legend title
  labs(title = "What is the most avarage country in Europe?",
       subtitle = "Countries closest to the avarage European levels of gdp per capita, social ineq
       caption = "Source: GDP per capita from 'World Bank', Gini Index after tax from 'Our World i
  theme_fivethirtyeight()+   #Pleasent them
  theme(panel.grid.major = element_blank(),   #Removing longditude and latitude
        axis.text = element_blank(),   #No unnecessary titles
        plot.title = element_text(hjust = 0.5,
                                  size = 22,
                                  margin=margin(2,0,2,0)),
        plot.subtitle = element_text(hjust = 0.5,
                                     size = 15,
                                     margin=margin(10,0,10,0)),
        plot.caption = element_text(hjust = 0.5),
        legend.title = element_text(size = 15),
```

```
        legend.text = element_text(size = 12))+
  guides(fill = guide_legend(ncol = 4))

# Showing and saving map

ggsave("map5.png", map5, width = 11, height = 7)

img <- magick::image_read("map5.png")    #Plotting the raw png file to R console
plot(img)
```



Central Europe seems to be the most average European region. Countries like Poland, Germany and Austria are pretty rich, but not too much compared to their northern neighbours. They also don't score the best in terms of democracy or equality, which closes the gap to the average. Eastern Europe fare rather poorly, in part because they are much more authoritarian and poorer than the average. But there are exceptions, Armenia and Estonia for example, have succeeded in closing the gap to western Europe in a greater degree than their respective neighbours.