QMSS G4603: Data Visualization - Assignment 1

Grace Kong (gyk2108)

2/14/2017

Memo and Plots

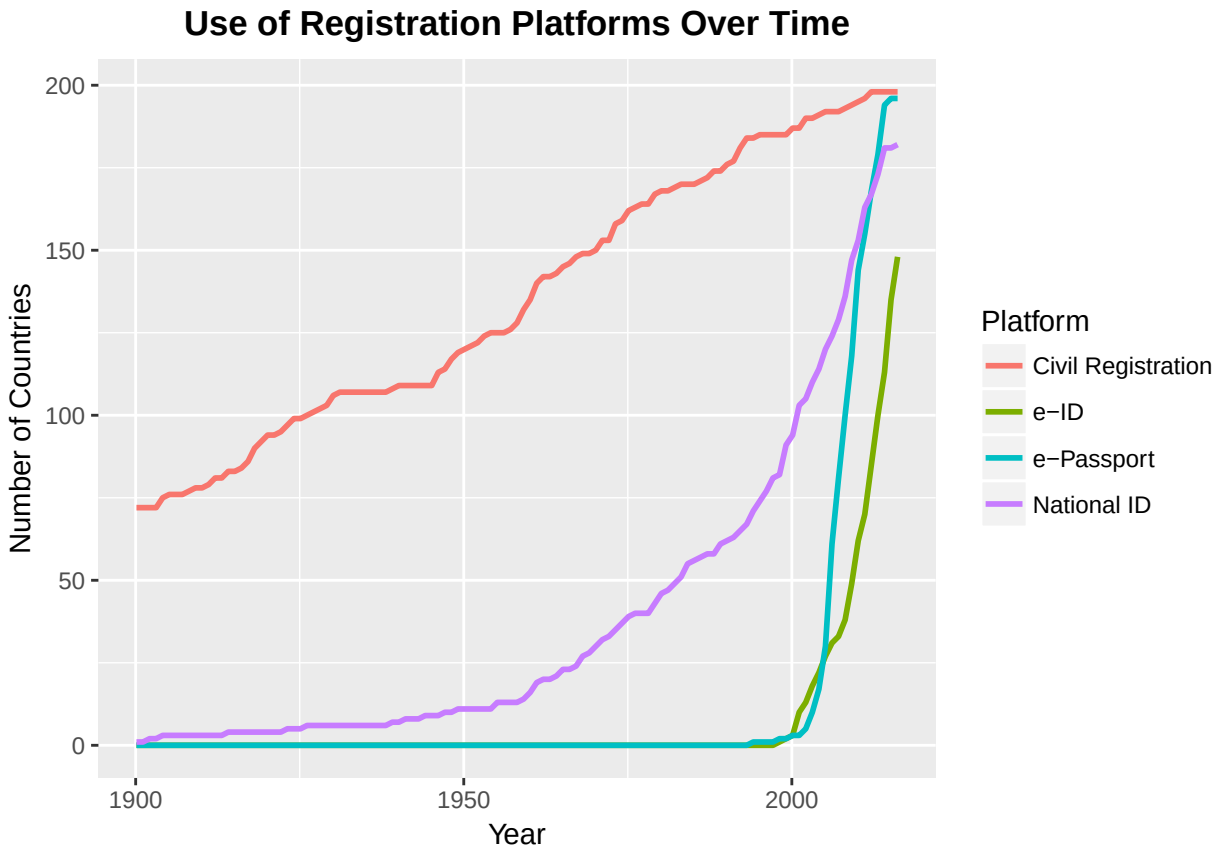
Introduction

The World Bank's ID For Development (ID4D) dataset contains information on four platforms for government registration of its citizen population: birth (or civil) registration, national IDs, e-IDs and e-Passports. A country typically aims to register its as much of its population as possible, that is, aim for a registration rate close to 100%, for the purposes of information and planning, conducting census and elections, national security, and so forth.

Use and Associated Policies of Registration Platforms Worldwide

Figure 1 plots how the number of countries using these four platforms has trended over time. Civil or birth registration has historically been the most common platform, already in use in 72 countries in 1900. National ID registration increased in popularity over the 20th Century, picking up especially from the 1950s onward. With the recent availability of e-technologies, the use of e-IDs and e-Passports began in 1998 and 1994 respectively and has increased steeply over the past two decades. They are presently used by 148 and 196 countries respectively, and the use of e-Passports even exceeds that of national IDs. Today, all four platforms are used by the majority of the 198 countries (that were included in the dataset).

Figure 1:



In the further analyses, we will zoom in on the use of birth registration and national ID registration, examining how successful governments have been in registering their populations as measured by registration rates.

Figure 2 takes a broad look at the current policies in the management of national registration. Specifically, we note how many countries make registration mandatory and free of charge, which in theory makes it more likely for one to get registered. While (all) 198 countries use birth registration, it is mandatory in186 countries, and free of charge in 153 countries. Of the 182 countries using civil registration, 152 make it mandatory, while 104 make it free. Thus, compared to civil registration, national ID registration is less likely to be mandatory free of charge. Thus, this may be a contributing factor to higher birth registration rates than national ID registration rates on the whole.

Figure 2:

Relationship of Registration Rates to Demographic and Political Factors

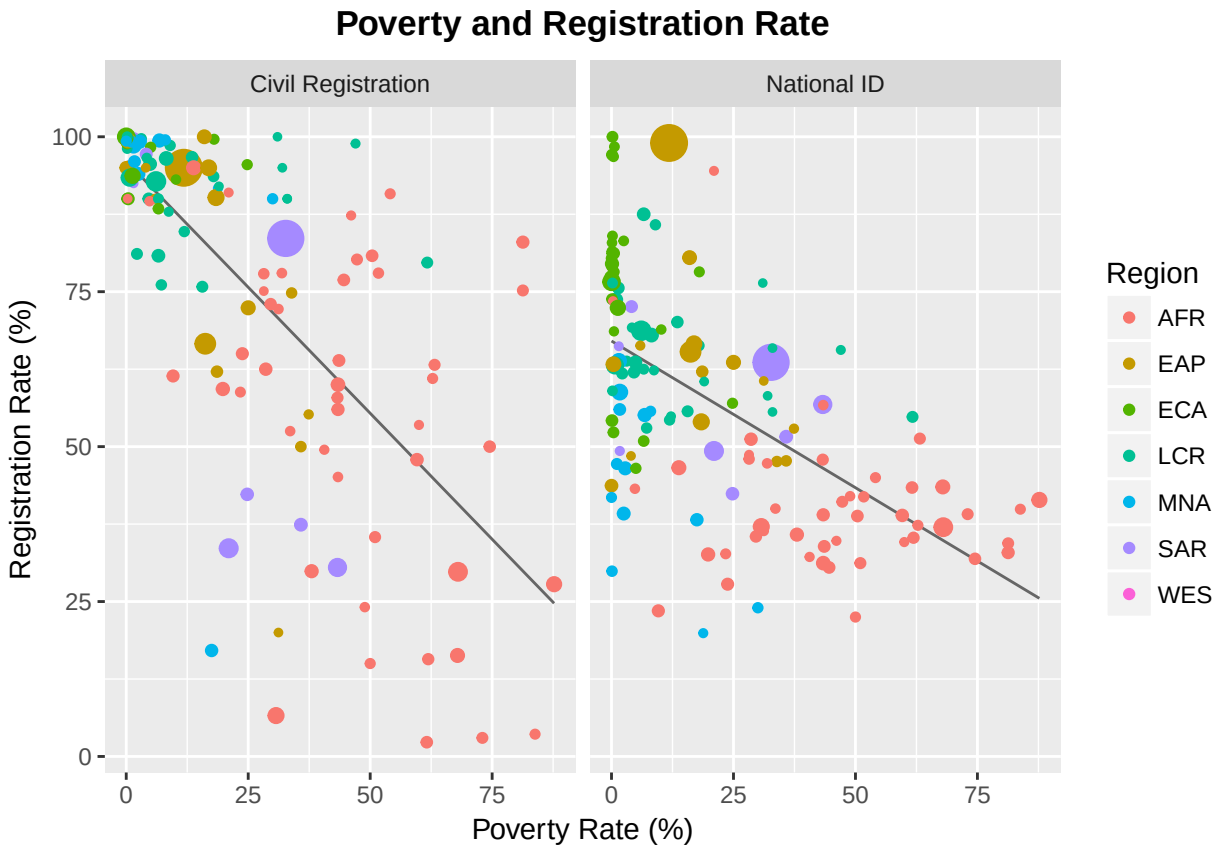
Next, Figures 3 to 7 examines the birth and national ID registration rates of countries and how they vary with various demographic and political factors.

Figure 3 analyzes the relationship between population and registration success. While we might hypothesize that massive countries may face more challenges registering their populations, we observe in the data that there is no clear upward or downward trend in birth / national ID registration rates as population increases. No unambiguous association exists when limiting to any of the income statuses either. Thus, it seems that registering a large population can be successfully managed with effective institutional systems in place, or there may perhaps be economies of scale in registering populations.

Figure 3:

Figure 4 looks at how registration rates vary with the wealth of a country. This time there is a clear negative correlation between poverty rates and registration rates. Poorer countries tend to be less successful in registering their populations, with a stronger correlation for birth registration than for national ID registration. We also see detail of the region (color) and population (size) of each country. We observe that Eastern Europe and Central Asia (ECA) and Western countries (WES) have been the most successful in registering, while African countries (AFR) have been the least successful. We will take a closer look in Figure 5. A notable outlier in national ID registration success is China, who with a 99% registration rate greatly outperforms the rate predicted by its low income.

Figure 4:



Next, **Figure 5** makes a more detailed comparison of the regional success in birth and national ID registration across the geographical regions, displaying the summary statistics for each region by a boxplot. We can see that ECA and WES are the two most successful regions in both birth and national ID registration. The remaining regions are ranked similarly for birth and national ID registration success, with one notable exception. While Middle East and North Africa (MNA) performs near the top, almost on par with WES for birth registration, they are the second worst performing region in national ID registration, with a median registration rate below 50%. For both platforms, AFR is the worst performing region.

Figure 5:

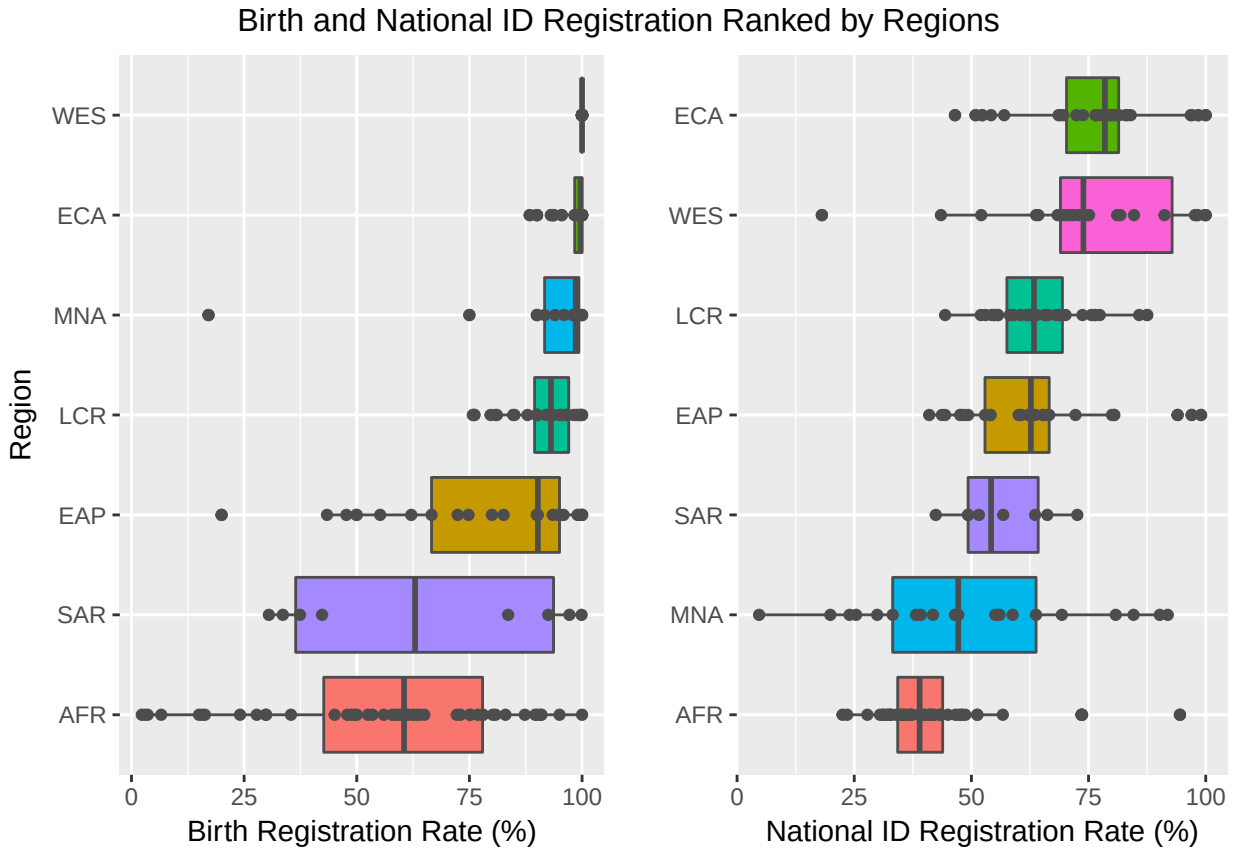
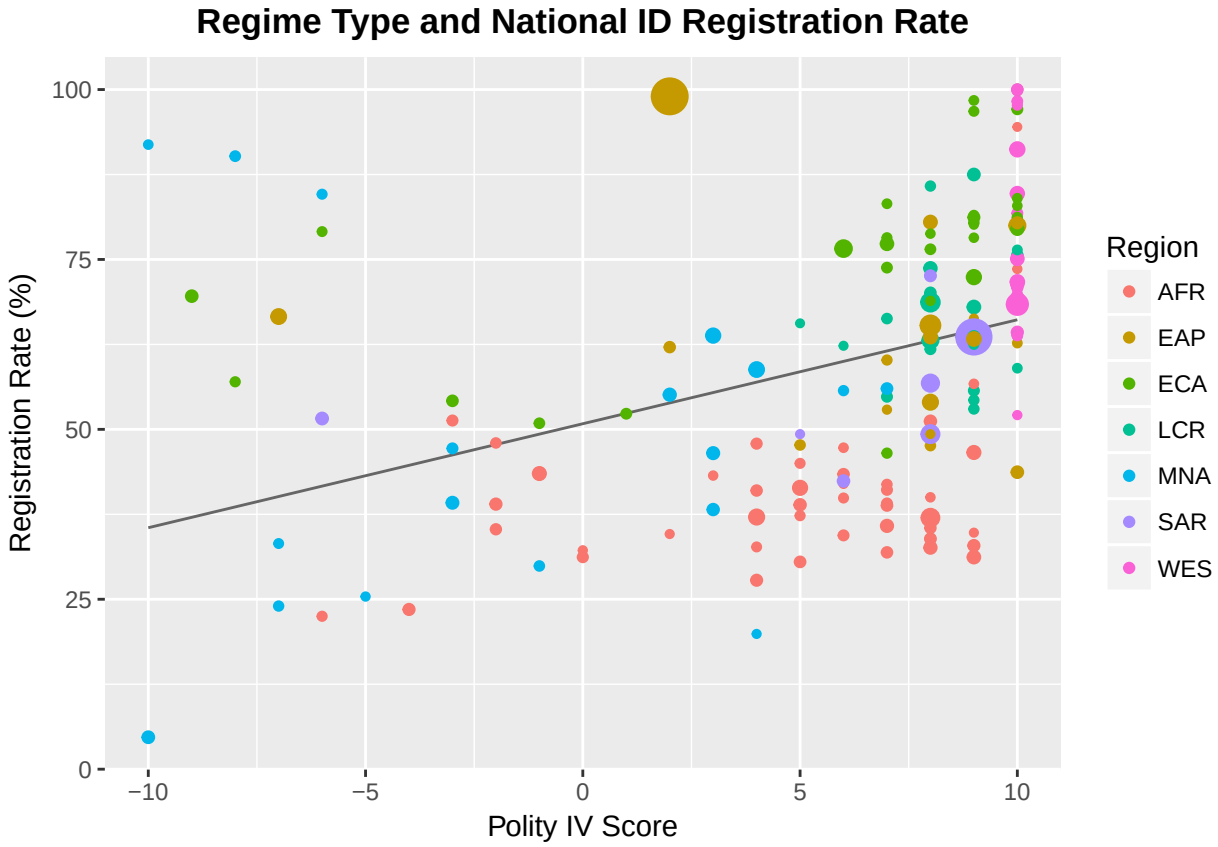


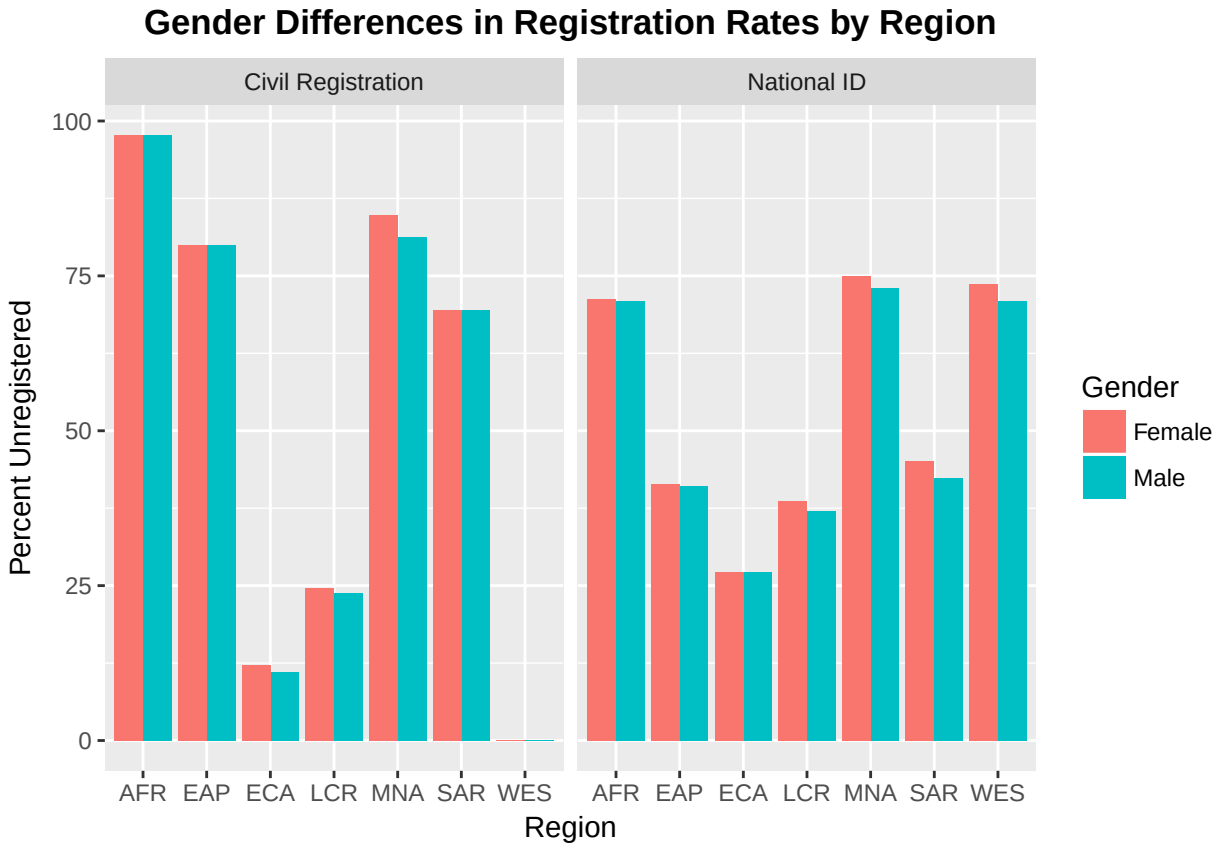
Figure 6 analyzes the relationship between regime type and national ID registration rate. A higher Polity IV score is used as a measure that a country is more like a democracy (positive scores) than a dictatorship (negative scores. It finds a positive correlation between national ID registration rate and being democratic. Some exceptions, however, are that some Middle Eastern countries (Oman, Qatar and UAE) that are strong dictatorships with high national ID registration, while numerous African countries are democracies with low national ID registration. Notably, China outperforms the trend as well, with a middle Polity IV score but very high national ID registration.

Figure 6:



Finally, **Figure 7** displays gender differences in birth and national ID registration rates, showing the percent unregistered. For birth registration, there are slightly more unregistered females in Middle East and North Africa. For national ID registration, there are slightly more unregistered females in South Asia and Western countries. For the others, the genders are roughly on par.

Figure 7:



Code and Description of Thought Process

Code - General Data Cleaning

```
rm(list = ls())

library(readr)
library(dplyr)
library(ggplot2)
require(gridExtra)

## Import data
raw_data <- read_csv("/Users/gracekongyx/Documents/*6_Data Science/QMSS G4063 Data Visualization/Assignm


## Rename variables

names(raw_data) <- raw_data[3, ]

# Retain selected variables, subset to proper observations
countries <- raw_data[4:201, c(3, 4, 5, 6, 7, 10, 11, 12, 14,
    17, 18, 20, 21, 22, 23, 24, 25, 31, 32, 34, 45, 46, 47, 48,
    49, 61, 62, 63, 65, 66, 85, 86, 87, 88, 89, 92, 93, 94, 95,
    99, 101, 102, 103, 104, 108, 109, 110, 113, 114, 115, 116,
    129, 137, 138, 188, 191, 192, 196, 197, 201, 208, 209, 214,
    215, 227, 228)]

names(countries) <- gsub(" |-", "_", names(countries))
names(countries) <- gsub("%", "_pct", names(countries))
names(countries) <- gsub("__", "_", names(countries))

names(countries)[c(48, 49, 50, 51, 52)] <- c("NID_unreg_M_pct",
    "NID_unreg_F_pct", "Unreg_15over", "U_15over_F_pct", "Pov_pct_p")
names(countries)[c(4, 5, 17, 43)] <- c("Code2", "Code3", "NID_M2",
    "Population_v")


## Modify values for some variables

# Standardize, consider all costs that aren't 'free' as 'not
# free' (due to the varied text representation)
countries$CR_Cost[countries$CR_Cost != "free"] <- "Not Free"
countries$NID_Cost[countries$NID_Cost != "free"] <- "Not Free"
countries$CR_Cost[countries$CR_Cost == "free"] <- "Free"
countries$NID_Cost[countries$NID_Cost == "free"] <- "Free"

# Notice that many, but not all, of the countries with Levels
# as NA are actually Western countries - from North America
# (excluding Mexico), Western Europe, and Australia and New
# Zealand Therefore code 'WES' as its separate region (not
# this is not WB terminology) In further regional analyses,
# the remaining NA countries (6 of them) will be excluded
countries$Region[countries$Economy %in% c("Australia", "Austria",
```

```
    "Belgium", "Canada", "Cyprus", "Denmark", "Finland", "France",
    "Germany", "Greece", "Iceland", "Ireland", "Italy", "Luxembourg",
    "Monaco", "Netherlands", "New Zealand", "Norway", "Portugal",
    "Spain", "Sweden", "Switzerland", "United Kingdom", "United States of America")] <- "WES"


## Change variable type

# Numeric variables
for (j in c(14, 24:25, 28:43, 45:53, 57, 59:66)) {
    countries[[j]] <- gsub(",", "", countries[[j]])
    countries[[j]] <- as.numeric(countries[[j]])
}

# Date variables (Year)
for (j in c(7, 11, 18, 23)) {
    countries[[j]] <- as.Date(strptime(countries[[j]], format = "%Y"))
}

# Factor variables
for (j in c(6, 8, 10, 12, 13, 16, 17, 19, 21, 22, 26, 27)) {
    countries[[j]] <- as.factor(countries[[j]])
}
levels(countries$CR_Org) <- c("Min Justice", "Min Interior",
    "Min Health", "Electoral", "Autonomous/Municipal", "6")
levels(countries$CR_M) <- c("Optional", "Mandatory")
levels(countries$NID_Org) <- c("Min of Justice", "Min of Interior",
    "Min of Health", "Electoral", "Autonomous/Municipal", "6")
levels(countries$NID_B) <- c("No NatID", "Not at birth", "Issued at birth")
levels(countries$NID_M) <- c("No NatID", "Optional", "Mandatory")
levels(countries$NID) <- c("No NatID", "NatID exists")
levels(countries$NID_M2) <- c("No", "Yes")
levels(countries$e_ID) <- c("No eID", "Identification only",
    "Some e-services", "Most e-services")
levels(countries$e_P) <- c("0", "1", "2")
levels(countries$e_P_Org) <- c("Min Justice", "Min Interior",
    "Min Health", "Electoral", "Autonomous/Municipal", "6")
levels(countries$DPL) <- c("No DPL", "Prepared", "Approved no agency",
    "Approved with agency")
levels(countries$RTI) <- c("No RTI Law", "Approved")

# Reorder levels for ordered factor variables
countries$Level <- factor(countries$Level, levels = c("LIC",
    "LMIC", "UMIC", "HIC"))


## Create new variables

# For standardization, want to compare the unregistered
# total, male and female populations at some point Hence
# convert from % registered to % registered
countries$BR_unreg_tot_pct <- 100 - countries$Birth_Reg_pct
countries$NID_unreg_tot_pct <- 100 - countries$Reg_Pop_pct
```

```
countries$BR_unreg_M_pct <- 100 - countries$BR_M_pct
countries$BR_unreg_F_pct <- 100 - countries$BR_F_pct
```

Code - Reshaping Datasets

In general, I performed several 'reshape longs', so as to later allow ggplot to vary on + Birth rate (BR) vs. national ID as a dependent variable (in order to combine plots using facets) + The unregistered percentage of males vs. females (in order to plot it in a combined graph)

In the first reshape long, I created one set of observations for birth registration (BR) statistics and another set of observations for national ID (NID) statistics. The relevant statistics are: + % Registered + % Unregistered + % Male unregistered + % Female unregistered + Whether it is mandatory + Whether it is free (cost)

```
## Reshape long (1)

countries_long <- countries[, c("Economy", "Region", "Population", "Level",
    "Pov_pct_p", "Pol_IV", "Birth_Reg_pct", "Reg_Pop_pct", "BR_unreg_tot_pct",
    "NID_unreg_tot_pct", "BR_unreg_M_pct", "NID_unreg_M_pct", "BR_unreg_F_pct",
    "NID_unreg_F_pct", "CR_M", "CR_Cost", "NID_M", "NID_Cost")]

countries_long <- reshape(countries_long, varying = list(c("Birth_Reg_pct",
    "Reg_Pop_pct"), c("BR_unreg_tot_pct", "NID_unreg_tot_pct"), c("BR_unreg_M_pct",
    "NID_unreg_M_pct"), c("BR_unreg_F_pct", "NID_unreg_F_pct"), c("CR_M", "NID_M"),
    c("CR_Cost", "NID_Cost")), v.names = c("Reg_Pct", "Unreg_Tot_Pct", "Unreg_M_Pct",
    "Unreg_F_Pct", "Mandatory", "Cost"), timevar = "Reg_Type", times = c("Civil Registration",
    "National ID"), new.row.names = 1:1000, direction = "long")
countries_long$id <- NULL

# Reorder factor variable 'Mandatory''
countries_long$Mandatory <- as.character(countries_long$Mandatory)
countries_long$Mandatory[countries_long$Mandatory == "No NatID"] <- "None"
countries_long$Mandatory <- factor(countries_long$Mandatory, levels = c("None",
    "Optional", "Mandatory"))
countries_long$Cost[countries_long$Mandatory == "None"] <- NA
```

In the second reshape long, performed on the already 'long' dataset, I created one set of observations for each of unregistered total, male and female percentages in their respective popoulations.

```
## Reshape long (2)

names(countries_long)[names(countries_long) == "id"] <- "id_1"
countries_longer_gender <- reshape(countries_long, varying = c("Unreg_Tot_Pct",
    "Unreg_M_Pct", "Unreg_F_Pct"), v.names = c("Unreg_Pct"), timevar = "Gender",
    times = c("Total", "Male", "Female"), new.row.names = 1:5000, direction = "long")
countries_longer_gender$id <- NULL
```

Code - Creating Dataset for Timeline of Countries' Uptake of Various Registration Platforms

To do so, I created a sequence of year-dates going from 1900 to present (2016). We are also given the starting year for each platform type in each country.

For each year, for each of the four platform types - birth registration (BR), national ID registration (NID),

e-IDs and e-Passports, I summed up the total number of countries who had implemented the platform on or before the given year. This is the number of countries with effective platforms as of that year.

Then, for ease of plotting with ggplot later, I reshaped the dataset 'long' to vary the observations on the platform type.

```
## Uptake timeline
year_seq <- seq(as.Date("1900", format = "%Y"), as.Date("2016", format = "%Y"),
    by = "year")
uptake_timeline <- data.frame(Year = year_seq)

# Sum countries for Civil/Birth Registration (CR)
uptake_timeline$n_CR <- 0
for (i in 1:nrow(uptake_timeline)) {
    uptake_timeline$n_CR[i] <- sum(countries$CR_Yr <= uptake_timeline$Year[i],
        na.rm = TRUE)
}

# Sum countries for National ID (Nat ID)
uptake_timeline$n_NID <- 0
for (i in 1:nrow(uptake_timeline)) {
    uptake_timeline$n_NID[i] <- sum(countries$NID_Yr <= uptake_timeline$Year[i],
        na.rm = TRUE)
}

# Sum countries for e-ID
uptake_timeline$n_e_ID <- 0
for (i in 1:nrow(uptake_timeline)) {
    uptake_timeline$n_e_ID[i] <- sum(countries$e_ID_Yr <= uptake_timeline$Year[i],
        na.rm = TRUE)
}

# Sum countries for e-Passport (e_P)
uptake_timeline$n_e_P <- 0
for (i in 1:nrow(uptake_timeline)) {
    uptake_timeline$n_e_P[i] <- sum(countries$e_P_Yr <= uptake_timeline$Year[i],
        na.rm = TRUE)
}

# Reshape dataset, varying on platform type
uptake_timeline <- reshape(uptake_timeline, varying = c("n_CR", "n_NID", "n_e_ID",
    "n_e_P"), v.names = "Number", timevar = "Reg_Type", times = c("Civil Registration",
    "National ID", "e-ID", "e-Passport"), new.row.names = 1:1000, direction = "long")
```

Code - Data Exploration and Refinement of Plots

I wanted to show how the number of countries using each platform changed over time, to capture which were the most popular platforms at each time period, and which period the uptake of each platform grew the most. Thus, I plotted frequency of use over time (after the data reshape).

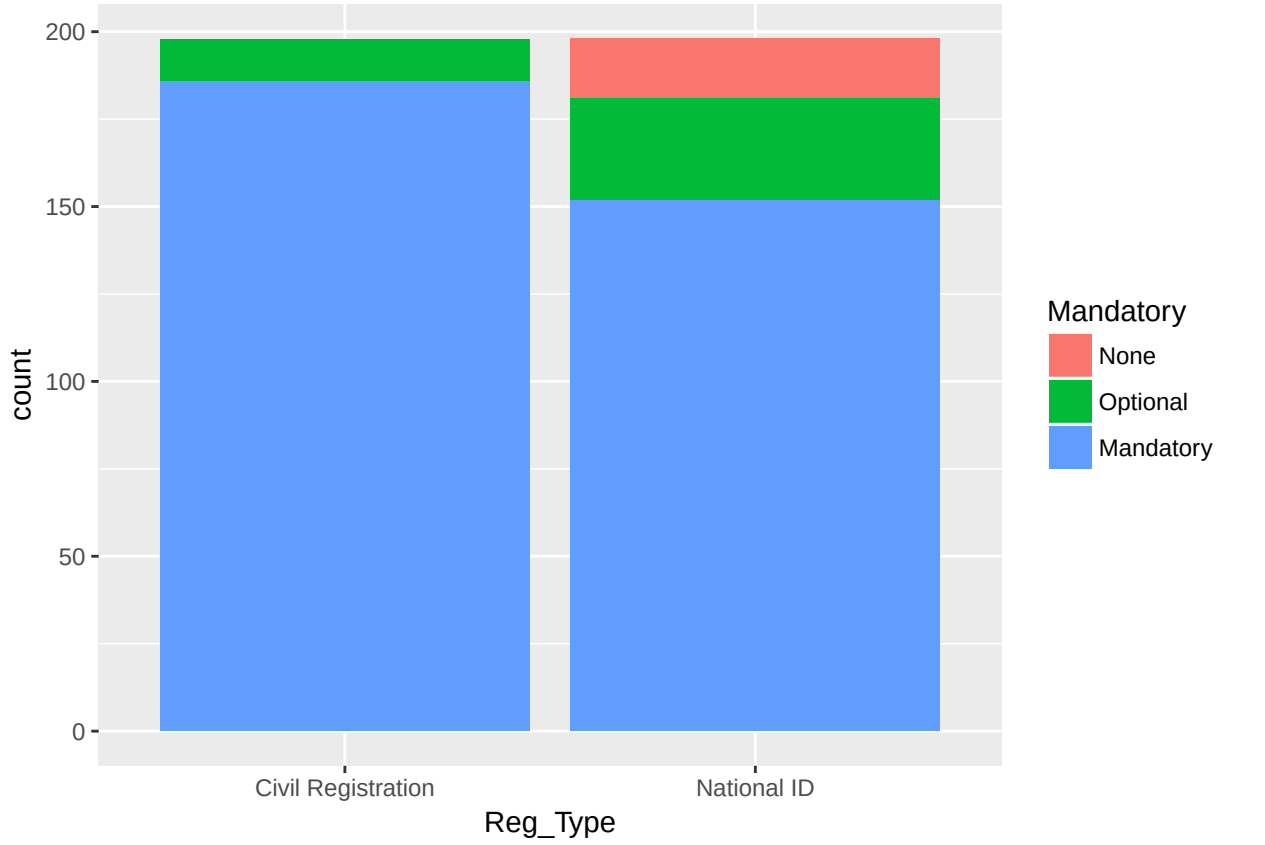
The following code gave rise to **Figure 1**.

```
# POLISHED PLOT 1
ggplot(uptake_timeline, aes(Year, Number, color = Reg_Type)) + geom_line(lwd = 1) +
    labs(x = "Year", y = "Number of Countries", color = "Platform") + ggtitle("Use of Registration Plat
```

```
    theme(plot.title = element_text(face = "bold", hjust = 0.5))
```

I tried plotting two bar charts side by side, showing the mandatory status for birth registration and NID through the colored fill.

```
# Process plot
ggplot(countries_long, aes(Reg_Type, fill = Mandatory)) + geom_bar()
```



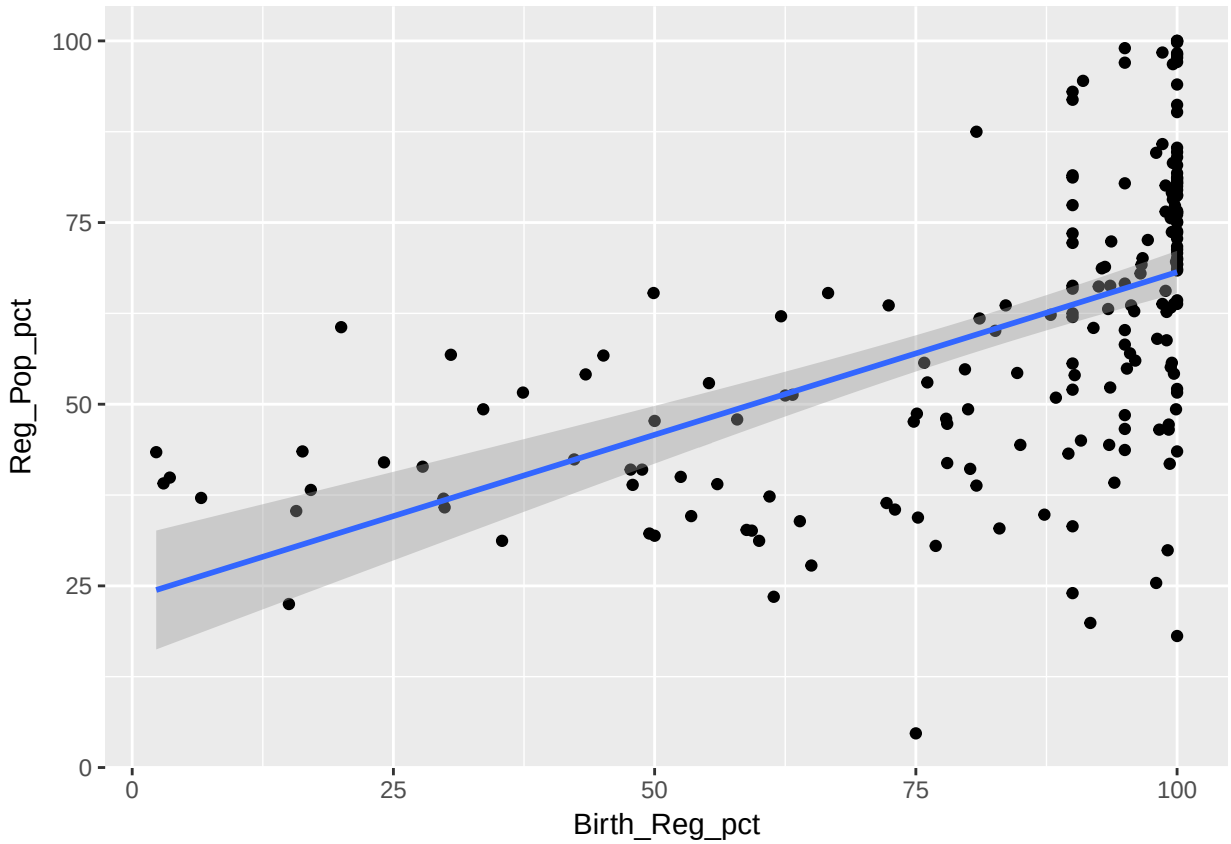
However, I decided it owuld be informative to capture another dimension, the cost of registration. Therefore, I moved birth registration vs. NID to the facets, used adjacent bars for optional vs. mandatory, and then segmented each bar by the cost status.

The following code gave rise to **Figure 2**.

```
# POLISHED PLOT 2
ggplot(countries_long, aes(Mandatory, fill = Cost)) + geom_bar() + labs(x = " ",
    y = "Number of Countries", fill = "Cost of \nRegistration") + facet_grid(. ~
    Reg_Type) + ggtitle("Registration Policies and Costs", ) + theme(plot.title = element_text(face = "
    hjust = 0.5))
```

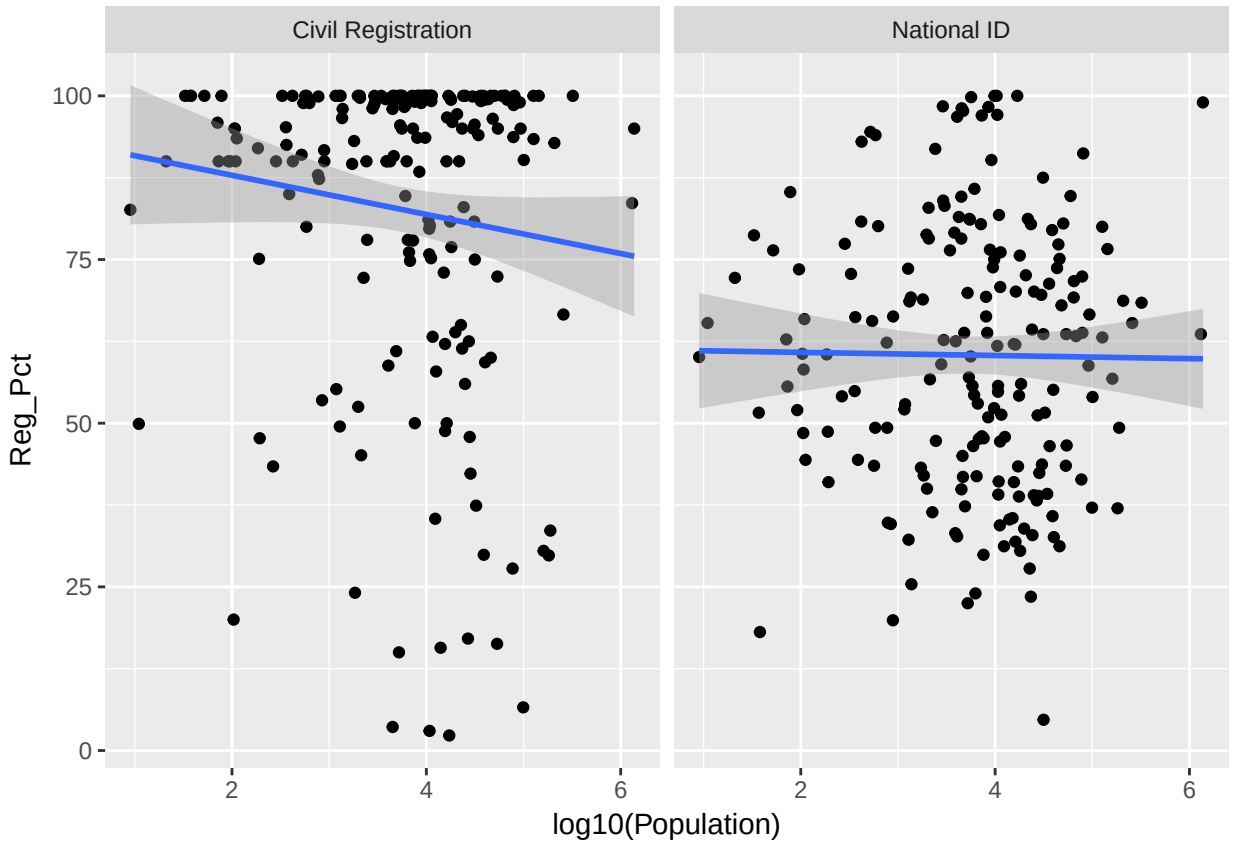
Next, I wanted to select relevant dependent variables, and shortlisted birth registration percentage and national ID registration percentage (reg_pop_pct). To see if there was a difference in using either of the two, I ran a scatter between the two. I found a positive, but not very strong, correlation between the two, noticing that many countries tended to have higher birth registration than national ID registration. Therefore, this implied that I should plot these two dependent variables separately (and not do away with one of them).

```
# Process plot
ggplot(countries, aes(Birth_Reg_pct, Reg_Pop_pct)) + geom_point() + geom_smooth(method = "lm")
```



Next is a preliminary graph of how civil and national ID registration relates to population size. I found very weak or no correlations with population size (logged).

```
# Process plot
ggplot(countries_long, aes(log10(Population), Reg_Pct)) + geom_point() + geom_smooth(method = "lm") +
    facet_grid(. ~ Reg_Type)
```




```
# ggplot(countries_long, aes(log10(Population), Birth_Reg_pct)) +
# geom_point() + scale_x_log10() ggplot(countries, aes(log10(Population),
# Reg_Pop_pct)) + geom_point() + scale_x_log10()
```

As the lack of correlation is interesting in itself, I wanted to see if there was still no correlation between population and registration when restricting to any of the income levels. Therefore, I added mapped region onto income levels, and still found no correlation.

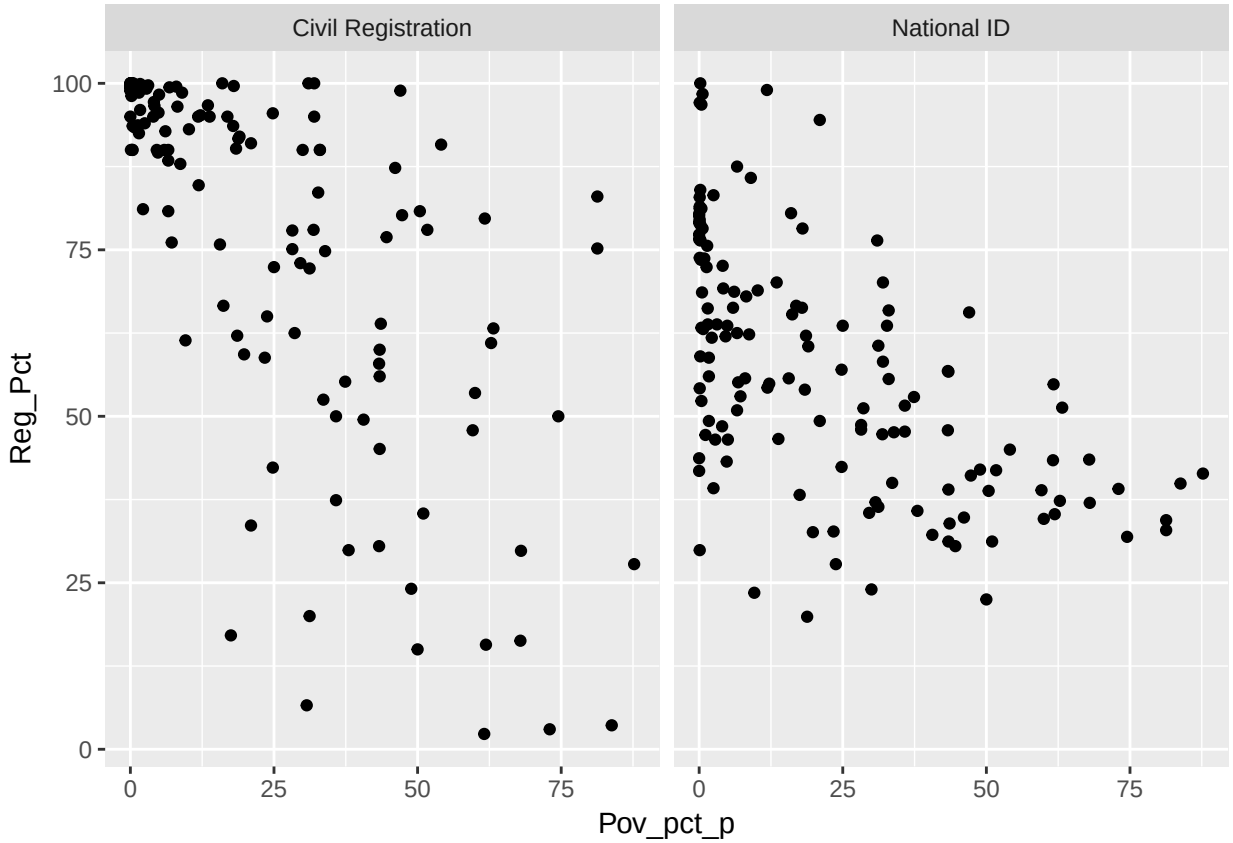
Thus, the following code gave rise to **Figure 3**.

```
# POLISHED PLOT 3
ggplot(countries_long, aes(log10(Population), Reg_Pct, color = Level)) + geom_point() +
    geom_smooth(se = FALSE) + labs(x = "Total Population (Logged)", y = "Registration Rate (%)",
    color = "Income \nLevel") + scale_x_continuous(breaks = 1:6) + facet_grid(. ~
    Reg_Type) + ggtitle("Population and Registration Rate", ) + theme(plot.title = element_text(face =
    hjust = 0.5))
# ggplot(countries, aes(log10(Population), Birth_Reg_pct, color = Level)) +
# geom_point() ggplot(countries, aes(log10(Population), Reg_Pop_pct, color =
# Level)) + geom_point() + geom_smooth(se = FALSE)
```

However, what I did find was that looking at the position of the colored dots, there was a correlation between income level and registration rates. Thus, I investigated that next. This is a preliminary graph for the relationship between registration and wealth, where wealth is now measured by the poverty rate. The initial scatters show a clear negative relationship between poverty and registration rate.

```
# Process plot
ggplot(countries_long, aes(Pov_pct_p, Reg_Pct)) + geom_point() + facet_grid(. ~
    Reg_Type)
```

```
## Warning: Removed 114 rows containing missing values (geom_point).
```



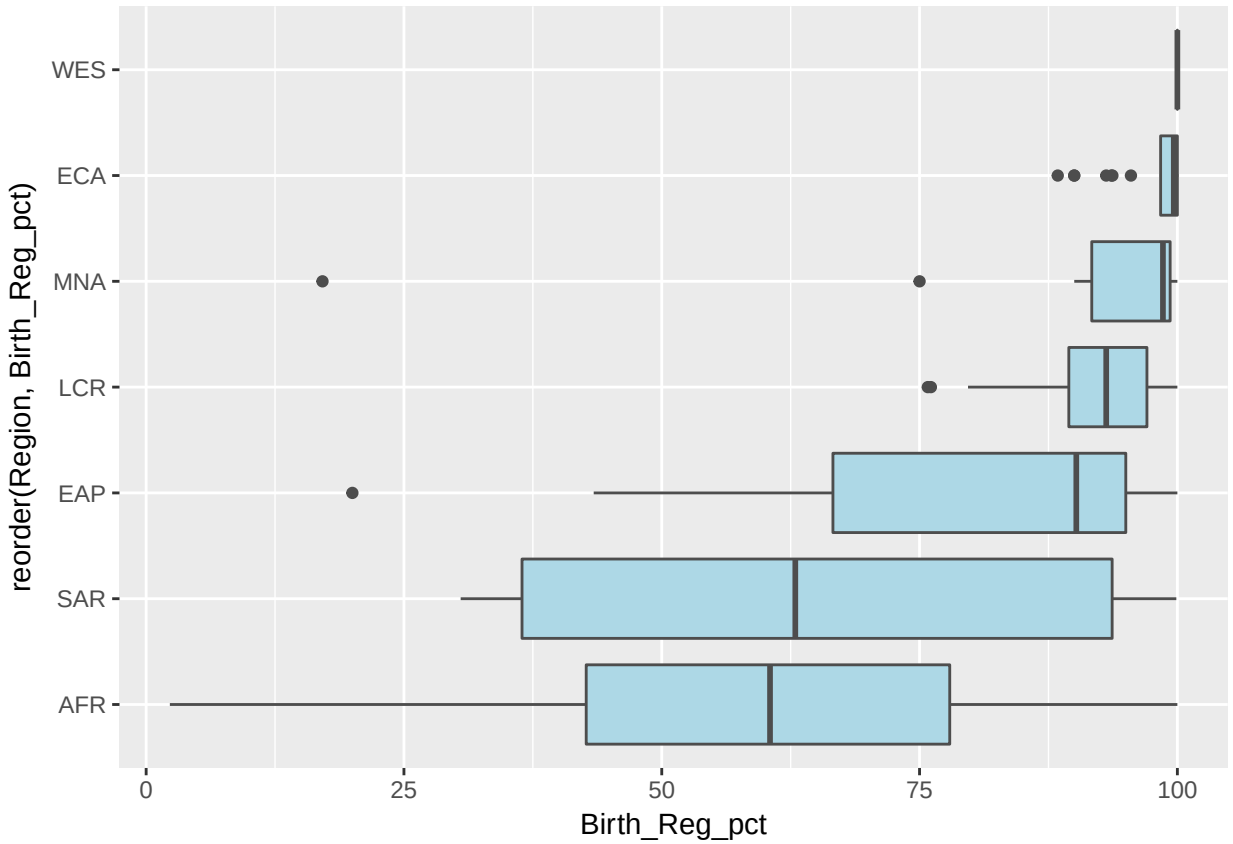

```
# ggplot(countries, aes(Pov_pct_p, Birth_Reg_pct)) + geom_point()
# ggplot(countries, aes(Pov_pct_p, Reg_Pop_pct)) + geom_point()
```

Therefore, I further mapped geographical region onto color and mapped population onto size of the dots. For the new variable introduced, region, I noticed a correlation between that and registration. The following code gave rise to **Figure 4**.

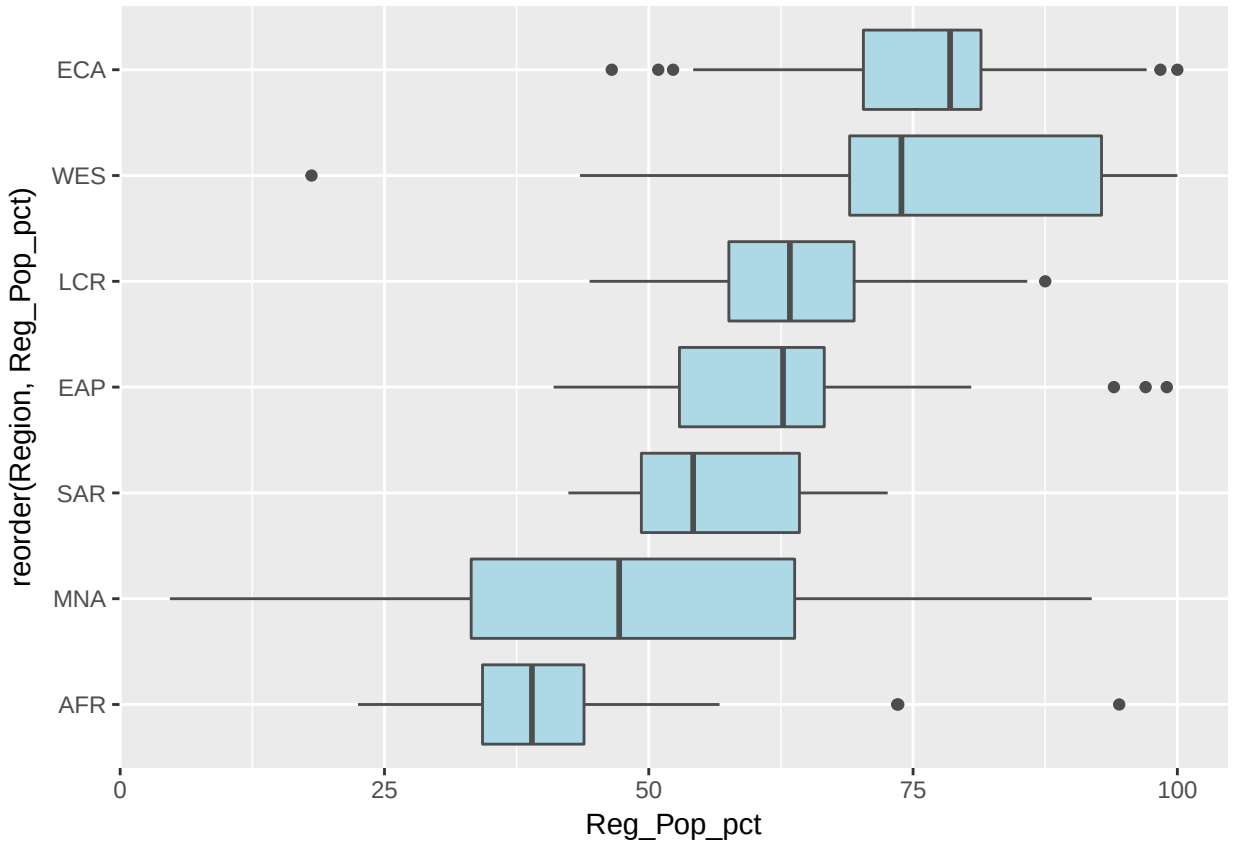
```
# POLISHED PLOT 4
ggplot(data = filter(countries_long, !is.na(Region)), aes(Pov_pct_p, Reg_Pct)) +
    geom_smooth(method = "lm", color = "gray40", lwd = 0.5, se = FALSE) + geom_point(aes(color = Region
    size = Population/10^6)) + labs(x = "Poverty Rate (%)", y = "Registration Rate (%)",
    color = "Region") + scale_size_continuous("Population", breaks = c(25, 50,
    75, 100)) + facet_grid(. ~ Reg_Type) + ggtitle("Poverty and Registration Rate",
    ) + theme(plot.title = element_text(face = "bold", hjust = 0.5))
```

Next, I investigated regional differences in registration, choosing a boxplot representation so that it could show the key summary statistics and dispersion of registration rates among countries in a region. I plotted birth registration and national ID registration separately at first. However, I noticed that the ranked order of the regions changed.

```
# Process plot
ggplot(data = filter(countries, !is.na(Region)), aes(reorder(Region, Birth_Reg_pct),
    Birth_Reg_pct)) + geom_boxplot(color = "gray30", fill = "light blue") +
    coord_flip()
```




```
# Process plot
ggplot(data = filter(countries, !is.na(Region)), aes(reorder(Region, Reg_Pop_pct),
    Reg_Pop_pct)) + geom_boxplot(color = "gray30", fill = "light blue") + coord_flip()
```

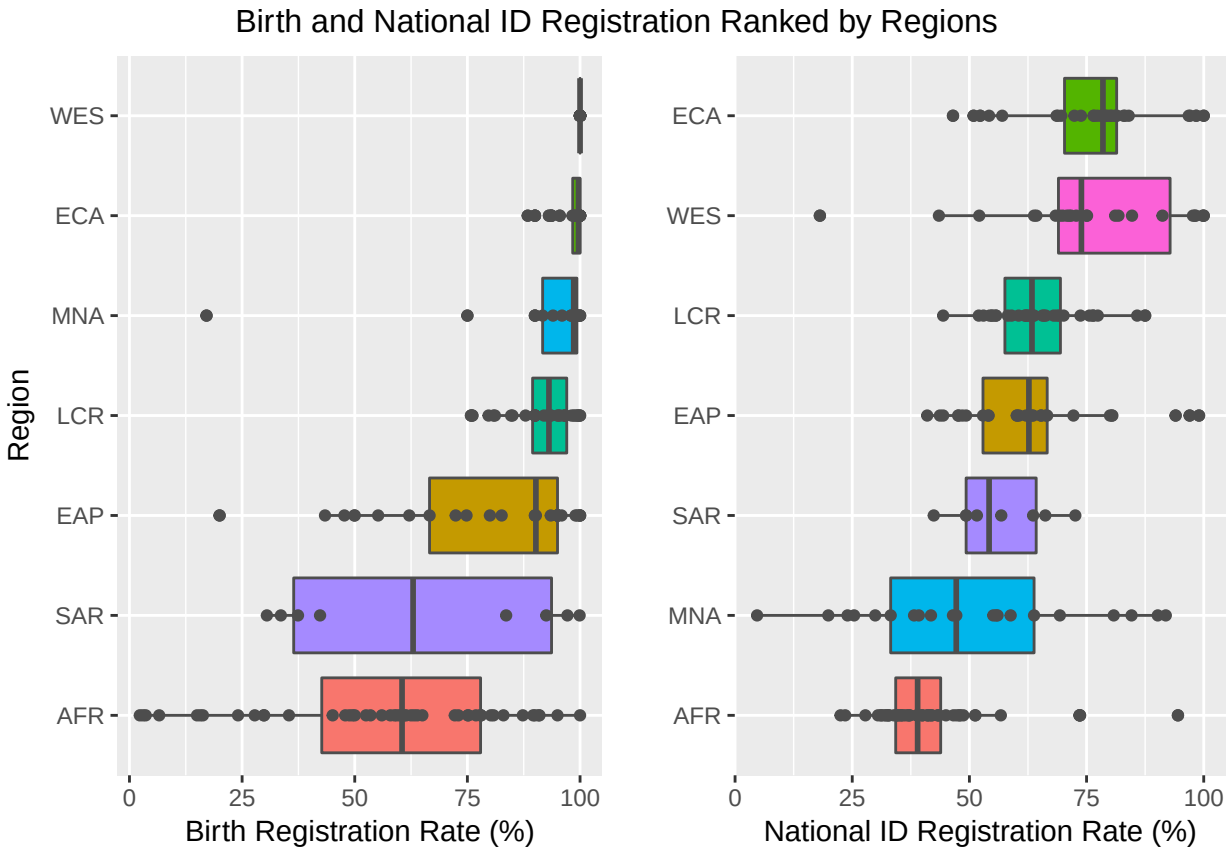


Therefore, it would be more effective to have these plots side by side. I also colored in the countries to aid comparison between the two charts. The following code gave rose to **Figure 5**.

```
plot1 <- ggplot(data = filter(countries, !is.na(Region)), aes(reorder(Region,
    Birth_Reg_pct), Birth_Reg_pct)) + geom_boxplot(aes(fill = Region), color = "gray30") +
    geom_point(color = "gray30") + labs(x = "Region", y = "Birth Registration Rate (%)") +
    coord_flip() + theme(legend.position = "none")

plot2 <- ggplot(data = filter(countries, !is.na(Region)), aes(reorder(Region,
    Reg_Pop_pct), Reg_Pop_pct)) + geom_boxplot(aes(fill = Region), color = "gray30") +
    geom_point(color = "gray30") + labs(x = " ", y = "National ID Registration Rate (%)") +
    coord_flip() + theme(legend.position = "none")

# POLISHED PLOT 5
grid.arrange(plot1, plot2, ncol = 2, top = "Birth and National ID Registration Ranked by Regions")
```

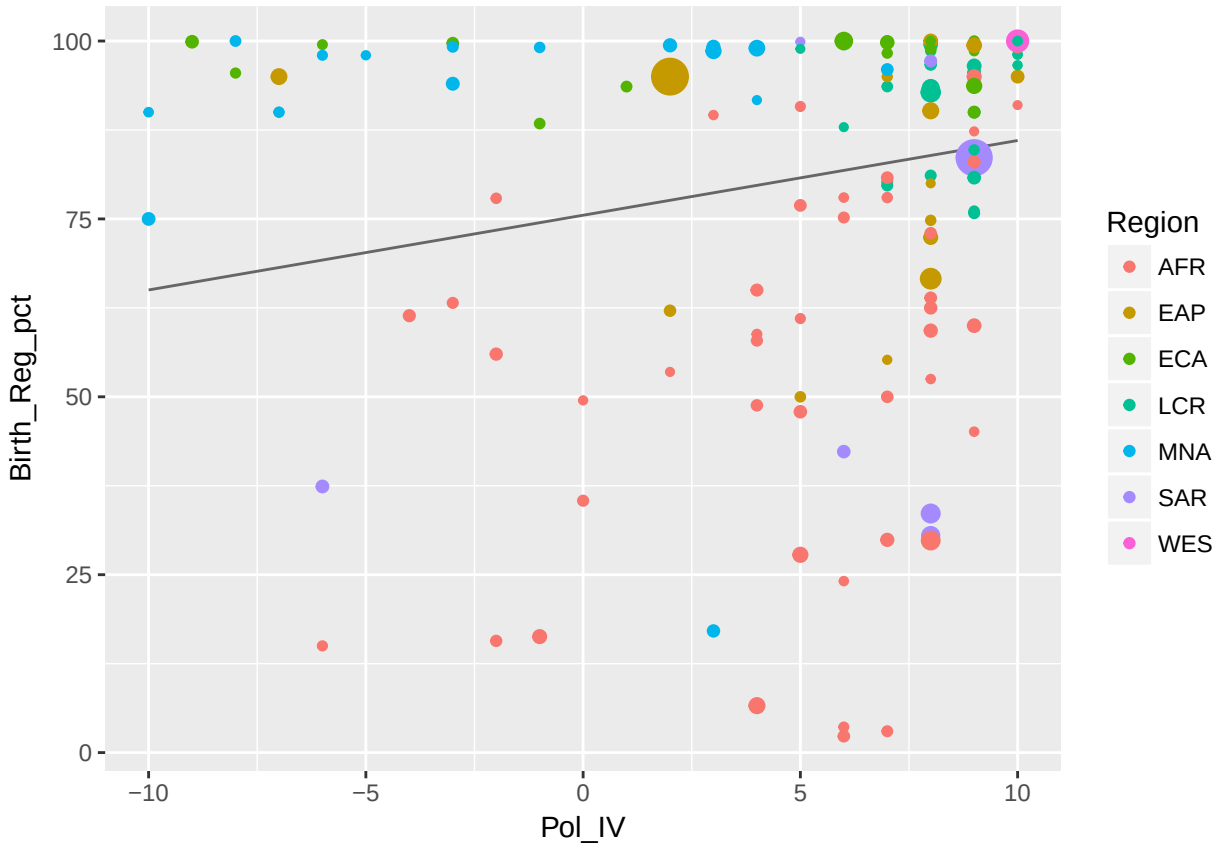


Next, I investigated the relationship between registration rates and regime type, as measured by polity score. I ran tha graphs for both birth registration and national ID registration, however decided that national ID registration was more relevant as it may aid the maintainence of voter rolls and such. Thus I discarded the following graph.

```
# Process plot
ggplot(data = filter(countries, !is.na(Region)), aes(Pol_IV, Birth_Reg_pct)) +
    scale_size_continuous(breaks = c(25, 50, 75, 100)) + geom_smooth(method = "lm",
    color = "gray40", lwd = 0.5, se = FALSE) + geom_point(aes(color = Region,
    size = Population/10^6))
```

```
## Warning: Removed 27 rows containing non-finite values (stat_smooth).
```

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



The following code gave rose to **Figure 6** on national ID registration and regime type.

```
# POLISHED PLOT 6
ggplot(data = filter(countries, !is.na(Region)), aes(Pol_IV, Reg_Pop_pct)) +
    geom_smooth(method = "lm", color = "gray40", lwd = 0.5, se = FALSE) + geom_point(aes(color = Region
    size = Population/10^6)) + labs(x = "Polity IV Score", y = "Registration Rate (%)",
    color = "Region") + scale_size_continuous(breaks = c(25, 50, 75, 100)) +
    ggtitle("Regime Type and National ID Registration Rate", ) + theme(plot.title = element_text(face =
    hjust = 0.5))
```

Finally, given that we had variables for the registration rates by gender, I was interested in finding out whether were gender differences, especially pertaining to particular geographical regions. Therefore I used the even-longer reshaped dataset which varied the observations by gender (see above explanation). This allowed me to plot gender side by side as 'dodged' column charts, across regions.

The following code gave rose to **Figure 7**.

```
# POLISHED PLOT 7
ggplot(data = filter(countries_longer_gender, Gender != "Total", !is.na(Region)),
    aes(Region, Unreg_Pct, fill = Gender)) + geom_col(position = "dodge") +
    labs(x = "Region", y = "Percent Unregistered", fill = "Gender") + facet_grid(. ~
    Reg_Type) + ggtitle("Gender Differences in Registration Rates by Region") +
    theme(plot.title = element_text(face = "bold", hjust = 0.5))
```