# Master's Thesis Data Cleaning

Diego Andrés Vásquez

21/03/2022

## 1. Load Libraries

```
library(tidyverse)
library(magrittr)
library(dplyr)
library(janitor)
library(readxl)
library(WDI)
```

```
## Warning: package 'WDI' was built under R version 4.0.5
```

```
library(stargazer)
```

## 2. Load Datasets

```
sfi_data <- read_xls("SFIv2018.xls")
prio_acd <- read_xlsx("ucdp-prio-acd-211.xlsx")
world_bank_data <- WDI(indicator = c("NY.GDP.MKTP.CD", "NY.GDP.MKTP.KD.ZG", "NY.GDP.PCAP.CD", "SI.POV.G
```

## 3. Clean Datasets: A) Adjust variable names + variable typology

```
str(world_bank_data)
```

```
## 'data.frame':    16226 obs. of  8 variables:
##  $ iso2c            : chr  "1A" "1A" "1A" "1A" ...
##  $ country          : chr  "Arab World" "Arab World" "Arab World" "Arab World" ...
##  $ year             : int  1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 ...
##  $ NY.GDP.MKTP.CD   : num  NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "GDP (current US$)"
##  $ NY.GDP.MKTP.KD.ZG: num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "GDP growth (annual %)"
##  $ NY.GDP.PCAP.CD   : num  NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "GDP per capita (current US$)"
##  $ SI.POV.GINI      : num  NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "Gini index (World Bank estimate)"
##  $ SP.POP.TOTL      : num  9.22e+07 9.47e+07 9.73e+07 1.00e+08 1.03e+08 ...
##   ..- attr(*, "label")= chr "Population, total"
```

```
names(world_bank_data)[names(world_bank_data) == "NY.GDP.MKTP.CD"] <- "gdp"
names(world_bank_data)[names(world_bank_data) == "NY.GDP.MKTP.KD.ZG"] <- "gdp_growth_rate"
names(world_bank_data)[names(world_bank_data) == "NY.GDP.PCAP.CD"] <- "gdp_p_cap"
names(world_bank_data)[names(world_bank_data) == "SI.POV.GINI"] <- "gini"
names(world_bank_data)[names(world_bank_data) == "SP.POP.TOTL"] <- "population"
names(prio_acd)[names(prio_acd) == "location"] <- "country"
prio_acd$year<- as.numeric(as.numeric(prio_acd$year))
world_bank_data <- world_bank_data %>% filter(year >= 1995, year <= 2018) #filter world bank data by ye
```

## 3. Clean Datasets: B) Manually adjust country names

```
#World Bank and SFI Adjustments
world_bank_data$country[world_bank_data$country=="Cabo Verde"] <- "Cape Verde"
world_bank_data$country[world_bank_data$country=="Egypt, Arab Rep."] <- "Egypt"
world_bank_data$country[world_bank_data$country=="Gambia, The"] <- "Gambia"
world_bank_data$country[world_bank_data$country=="Iran, Islamic Rep."] <- "Iran"
world_bank_data$country[world_bank_data$country=="Kyrgyz Republic"] <- "Kyrgyzstan"
world_bank_data$country[world_bank_data$country=="Lao PDR"] <- "Laos"
world_bank_data$country[world_bank_data$country=="North Macedonia"] <- "Macedonia"
world_bank_data$country[world_bank_data$country=="Timor-Leste"] <- "Timor Leste"
world_bank_data$country[world_bank_data$country=="Korea, Dem. People's Rep."] <- "Korea, North"
world_bank_data$country[world_bank_data$country=="Korea, Rep."] <- "Korea, South"
world_bank_data$country[world_bank_data$country=="Russian Federation"] <- "Russia"
world_bank_data$country[world_bank_data$country=="Eswatini"] <- "Swaziland"
world_bank_data$country[world_bank_data$country=="Syrian Arab Republic"] <- "Syria"
world_bank_data$country[world_bank_data$country=="Venezuela, RB"] <- "Venezuela"
world_bank_data$country[world_bank_data$country=="Yemen, Rep."] <- "Yemen"
world_bank_data$country[world_bank_data$country=="Congo, Dem. Rep."] <- "Dem. Rep. of Congo"
world_bank_data$country[world_bank_data$country=="Congo-Brazzaville."] <- "Congo-Brazzaville"
sfi_data$country[sfi_data$country=="Serbia & Montenegro"] <- "Serbia"
sfi_data$country[sfi_data$country=="Sudan (North)"] <- "Sudan"


#Prio Adjustments to World Bank and SFI
prio_acd$country[prio_acd$country=="India, Pakistan"] <- "India"
prio_acd$country[prio_acd$country=="Myanmar (Burma)"] <- "Myanmar"
prio_acd$country[prio_acd$country=="Yemen (North Yemen)"] <- "Yemen"
prio_acd$country[prio_acd$country=="DR Congo (Zaire)"] <- "Dem. Rep. of Congo"
prio_acd$country[prio_acd$country=="United States of America"] <- "United States"
prio_acd$country[prio_acd$country=="Ivory Coast"] <- "Cote d'Ivoire"
prio_acd$country[prio_acd$country=="North Macedonia"] <- "Macedonia"
prio_acd$country[prio_acd$country=="Russia (Soviet Union)"] <- "Russia"
prio_acd$country[prio_acd$country=="Cambodia (Kampuchea), Thailand"] <- "Thailand"
prio_acd$country[prio_acd$country=="Cambodia (Kampuchea)"] <- "Cambodia"
prio_acd$country[prio_acd$country=="Congo"] <- "Congo-Brazzaville"
prio_acd$country[prio_acd$country=="Serbia (Yugoslavia)"] <- "Serbia"
prio_acd$country[prio_acd$country=="Bosnia-Herzegovina"] <- "Bosnia and Herzegovina"
prio_acd$country[prio_acd$country=="Eritrea, Ethiopia"] <- "Eritrea"
prio_acd$country[prio_acd$country=="Cameroon, Nigeria"] <- "Cameroon"
prio_acd$country[prio_acd$country=="Djibouti, Eritrea"] <- "Djibouti"
prio_acd$country[prio_acd$country=="Afghanistan, United Kingdom, United States of America"] <- "Afghanis
```

```
prio_acd$country[prio_acd$country=="Afghanistan, United Kingdom, United States of America"] <- "Afghanis
prio_acd$country[prio_acd$country=="Iran, Israel"] <- "Iran"
prio_acd$country[prio_acd$country=="Ecuador, Peru"] <- "Ecuador"
prio_acd$country[prio_acd$country=="Australia, Iraq, United Kingdom, United States of America"] <- "Uni
prio_acd$country[prio_acd$country=="South Sudan, Sudan"] <- "South Sudan"

prio_acd$intensity_level[prio_acd$intensity_level==1] <- 0
prio_acd$intensity_level[prio_acd$intensity_level==2] <- 1
```

***NOTE*: Only Taiwan is unaccounted for between merge of SFI and World Bank Data because the World Bank does not have information on it.**

***NOTE*: In instances where conflict takes place in several countries, only the first country listed holds the conflict variable to not double count instances of war.**

## 4. Merge Datasets

```
step_1 <- full_join(sfi_data, world_bank_data, by=c("year", "country")) #merge data sets on year and co
step_2 <- full_join(step_1, prio_acd, by=c("year", "country")) #repeat above with third data set
#step_2 %>% select(country, year, type_of_conflict, gdp, population, gdp_growth_rate, gdp_p_cap, gini,
```

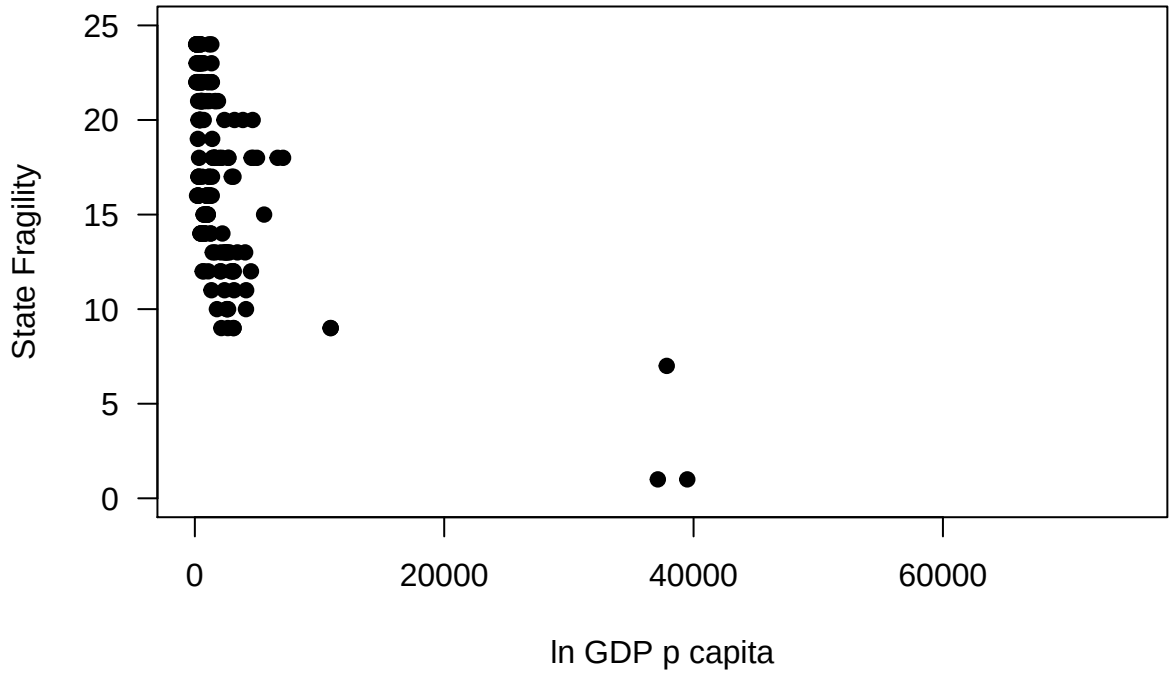
## 5. Simple Bivariate Regression

```
options(scipen=999)
model_1 <- lm(sfi ~ intensity_level*gdp_p_cap, data=step_2)
print(summary(model_1), digits = 2)
```

```
##
## Call:
## lm(formula = sfi ~ intensity_level * gdp_p_cap, data = step_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.71  -3.07   0.02   3.11   9.15
##
## Coefficients:
##                             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)                16.098617   0.156640   102.8 < 0.0000000000000002
## intensity_level1            2.319541   0.358329     6.5          0.0000000002
## gdp_p_cap                  -0.000353   0.000017   -20.3 < 0.0000000000000002
## intensity_level1:gdp_p_cap -0.000137   0.000061    -2.2                 0.03
##
## (Intercept)                ***
## intensity_level1           ***
## gdp_p_cap                  ***
## intensity_level1:gdp_p_cap *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.9 on 913 degrees of freedom
##   (7340 observations deleted due to missingness)
## Multiple R-squared:  0.37,   Adjusted R-squared:  0.37
## F-statistic: 1.8e+02 on 3 and 913 DF,  p-value: <0.0000000000000002
```

```
plot(sfi ~ gdp_p_cap, data=step_2, col=as.integer(intensity_level), xlab="ln GDP p capita",
     ylab="State Fragility", pch=19, las=1,
     xlim=c(0,75000))
```




## 6. Save Regression

```
stargazer(model_1,
          type = 'latex',
          covariate.labels = c("log GDP per capita", "ln Population", "Conflict Intensity"),
          dep.var.labels = "State Fragility")
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Tue, Mar 22, 2022 - 10:03:04
## \begin{table}[!htbp] \centering
##   \caption{}
```

```
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{1}{c}{\textit{Dependent variable:}} \\
## \cline{2-2}
## \\[-1.8ex] & State Fragility \\
## \hline \\[-1.8ex]
##  log GDP per capita & 2.320$^{***}$ \\
##   & (0.358) \\
##   & \\
##  ln Population & $-$0.0004$^{***}$ \\
##   & (0.00002) \\
##   & \\
##  Conflict Intensity & $-$0.0001$^{**}$ \\
##   & (0.0001) \\
##   & \\
##  Constant & 16.099$^{***}$ \\
##   & (0.157) \\
##   & \\
## \hline \\[-1.8ex]
## Observations & 917 \\
## R$^{2}$ & 0.373 \\
## Adjusted R$^{2}$ & 0.371 \\
## Residual Std. Error & 3.909 (df = 913) \\
## F Statistic & 181.357$^{***}$ (df = 3; 913) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{1}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```