Dat Cleaning: Interfamilial Violence and Corruption in Guatemala

Rachel Rosenberg

7/4/2020

Contents

Setting up the Data

The 2012 and 2013 data sets include municipal-level information on birth rates, crime rates, homicide rates, and interfamiliar violence (i.e. domestic violence).

To begin:

Set the working directory and clear the environment;

```
#Set the working directory
setwd("/Users/racheleryn/Documents/Git/Guatemala Project")

#Clear the environment
rm(list=ls(all=TRUE))
```

Upload the 2012 and 2013 data;

```
#Call the rio package
library(rio)

#Import the Excel documents
g_2012 = import("2012.xlsx") #the data from 2012 will be called g_2012
g_2013 = import("2013.xlsx") #the data from 2013 will be called g_2013
```

Preview the g_2012 data;

```
str(g_2012)
```

```
## 'data.frame':    334 obs. of  6 variables:
##  $ departamento           : chr  "Alta Verapaz" "Alta Verapaz" "Alta Verapaz" "Alta Verapaz" ...
##  $ municipio              : chr  "Cobán" "Santa Cruz Verapaz" "San Cristóbal Verapaz" "Tactic" ...
##  $ tasa bruta de natalidad: num  23.8 23.8 35.3 29.7 27.5 33.5 38.7 40.7 27.6 26.7 ...
##  $ tasa de criminalidad   : num  217.4 59.7 70.8 105.2 34.9 ...
##  $ tasa de homicidios     : num  20.6 8.1 13.2 25.6 5 4.8 8.4 4.5 3.6 3.5 ...
##  $ violencia intrafamiliar: num  4.1 2.1 3.3 3.6 3.3 1.6 2.5 1.7 1.4 2.3 ...
```

Preview the g_2013 data;

```
str(g_2013)
```

```
## 'data.frame':    354 obs. of  6 variables:
##  $ departamento           : chr  "Alta Verapaz" "Alta Verapaz" "Alta Verapaz" "Alta Verapaz" ...
##  $ municipio              : chr  "Cobán" "Santa Cruz Verapaz" "San Cristóbal Verapaz" "Tactic" ...
##  $ tasa bruta de natalidad: num  23.8 23.8 35.3 29.7 27.5 33.5 38.7 40.7 27.6 26.7 ...
##  $ tasa de criminalidad   : num  217.4 59.7 70.8 105.2 34.9 ...
##  $ tasa de homicidios     : num  20.6 8.1 13.2 25.6 5 4.8 8.4 4.5 3.6 3.5 ...
##  $ violencia intrafamiliar: num  4.1 2.1 3.3 3.6 3.3 1.6 2.5 1.7 1.4 2.3 ...
```

Section 1: Clean Data

1. Put the municipality and department variables in lowercase with the (tolower) command.

For 2012:

```
#Lower case for the municipality and department variables for g_2012 data
g_2012$municipio = tolower(g_2012$municipio)
g_2012$departamento = tolower(g_2012$departamento)

#Check that the changes went through
head(g_2012, n = 1)
```

```
##   departamento municipio tasa bruta de natalidad tasa de criminalidad
## 1 alta verapaz     cobán                   23.8                217.4
##   tasa de homicidios violencia intrafamiliar
## 1               20.6                     4.1
```

For 2013:

```
#Lowering case for the municipality and department variables for g_2013 data
g_2013$municipio = tolower(g_2013$municipio)
g_2013$departamento = tolower(g_2013$departamento)

#Check that the changes went through
head(g_2013, n = 1)
```

```
##   departamento municipio tasa bruta de natalidad tasa de criminalidad
## 1 alta verapaz     cobán                   23.8                217.4
##   tasa de homicidios violencia intrafamiliar
## 1               20.6                     4.1
```

2. Remove the accents from those same lowercase department and municipality variables.

```
#Create a function to remove accents
remove.accents = function(s){
  old1 = "áéóíúñ"
  new1 = "aeoiun"
  s1 = chartr(old1, new1, s)
}
```

For 2012:

```
#Remove accents for the g_2012 data
g_2012$departamento = remove.accents(g_2012$departamento)
g_2012$municipio = remove.accents(g_2012$municipio)

#Check that the changes went through
head(g_2012, n = 1)
```

```
##   departamento municipio tasa bruta de natalidad tasa de criminalidad
## 1 alta verapaz     coban                   23.8                217.4
##   tasa de homicidios violencia intrafamiliar
## 1               20.6                     4.1
```

For 2013:

```
#Remove accents for the g_2013 data
g_2013$departamento = remove.accents(g_2013$departamento)
g_2013$municipio = remove.accents(g_2013$municipio)

#Check that the changes went through
head(g_2013, n = 1)
```

```
##   departamento municipio tasa bruta de natalidad tasa de criminalidad
## 1 alta verapaz     coban                   23.8                217.4
##   tasa de homicidios violencia intrafamiliar
## 1               20.6                     4.1
```

3. Rename the variables so that there are no spaces and replace those spaces with underscores.

For 2012:

```
#Fix the column names for g_2012
colnames(g_2012) = gsub(" ", "_", colnames(g_2012))

#Check that the changes went through
head(g_2012, n = 1)
```

```
##   departamento municipio tasa_bruta_de_natalidad tasa_de_criminalidad
## 1 alta verapaz     coban                   23.8                217.4
##   tasa_de_homicidios violencia_intrafamiliar
## 1               20.6                     4.1
```

For 2013

```
#Fix the column names for g_2013
colnames(g_2013) = gsub(" ", "_", colnames(g_2013))

#Check that the changes went through
head(g_2013, n = 1)
```

```
##   departamento municipio tasa_bruta_de_natalidad tasa_de_criminalidad
## 1 alta verapaz     coban                    23.8                217.4
##   tasa_de_homicidios violencia_intrafamiliar
## 1               20.6                     4.1
```

4. Add year variables to each of the files.

```
#Create variables with the respective years
year12 = 2012
year13 = 2013

#Add the variables to the data frames using the dplyr package
g_2012 = mutate(g_2012,year12)
g_2013 = mutate(g_2013,year13)

#Rename the variables using the dplyr package
g_2012 = g_2012 %>% rename(year = year12)
g_2013 = g_2013 %>% rename(year = year13)

#Check that the changes went through
colnames(g_2012) #for 2012
```

```
## [1] "departamento"            "municipio"
## [3] "tasa_bruta_de_natalidad" "tasa_de_criminalidad"
## [5] "tasa_de_homicidios"      "violencia_intrafamiliar"
## [7] "year"
```

```
colnames(g_2013) #for 2013
```

```
## [1] "departamento"            "municipio"
## [3] "tasa_bruta_de_natalidad" "tasa_de_criminalidad"
## [5] "tasa_de_homicidios"      "violencia_intrafamiliar"
## [7] "year"
```

5. Ensure that all variables are the correct class.

```
#Check the structure of the g_2012 data set
str(g_2012)
```

```
## 'data.frame':    334 obs. of  7 variables:
##  $ departamento           : chr  "alta verapaz" "alta verapaz" "alta verapaz" "alta verapaz" ...
##  $ municipio              : chr  "coban" "santa cruz verapaz" "san cristobal verapaz" "tactic" ...
##  $ tasa_bruta_de_natalidad: num  23.8 23.8 35.3 29.7 27.5 33.5 38.7 40.7 27.6 26.7 ...
```

```
##  $ tasa_de_criminalidad   : num  217.4 59.7 70.8 105.2 34.9 ...
##  $ tasa_de_homicidios     : num  20.6 8.1 13.2 25.6 5 4.8 8.4 4.5 3.6 3.5 ...
##  $ violencia_intrafamiliar: num  4.1 2.1 3.3 3.6 3.3 1.6 2.5 1.7 1.4 2.3 ...
##  $ year                   : num  2012 2012 2012 2012 2012 ...
```

```
#Check the structure of the g_2013 data set
str(g_2013)
```

```
## 'data.frame':    354 obs. of  7 variables:
##  $ departamento           : chr  "alta verapaz" "alta verapaz" "alta verapaz" "alta verapaz" ...
##  $ municipio              : chr  "coban" "santa cruz verapaz" "san cristobal verapaz" "tactic" ...
##  $ tasa_bruta_de_natalidad: num  23.8 23.8 35.3 29.7 27.5 33.5 38.7 40.7 27.6 26.7 ...
##  $ tasa_de_criminalidad   : num  217.4 59.7 70.8 105.2 34.9 ...
##  $ tasa_de_homicidios     : num  20.6 8.1 13.2 25.6 5 4.8 8.4 4.5 3.6 3.5 ...
##  $ violencia_intrafamiliar: num  4.1 2.1 3.3 3.6 3.3 1.6 2.5 1.7 1.4 2.3 ...
##  $ year                   : num  2013 2013 2013 2013 2013 ...
```

6. Label all of the variables using the labelled package.

```
#Change Labels for the g_2012 data set with the labelled package
var_label(g_2012) <- list(`departamento` = "Department",
                          `municipio` = "Municipality",
                          `tasa_bruta_de_natalidad` = "Birth Rate",
                          `tasa_de_criminalidad` = "Crime Rate",
                          `tasa_de_homicidios` = "Homicide Rate",
                          `violencia_intrafamiliar` = "Domestic Violence",
                          `year` = "Year")

#Change labels for the g_2013 data set
var_label(g_2013) <- list(`departamento` = "Department",
                          `municipio` = "Municipality",
                          `tasa_bruta_de_natalidad` = "Birth Rate",
                          `tasa_de_criminalidad` = "Crime Rate",
                          `tasa_de_homicidios` = "Homicide Rate",
                          `violencia_intrafamiliar` = "Domestic Violence",
                          `year` = "Year")
```

7. Save each of the cleaned cross-sectional data sets as Stata data sets.

```
export(g_2012, file = "clean_g_2012.dta") #the clean data set for the g_2012 data
export(g_2013, file = "clean_g_2013.dta") #the clean data set for the g_2013 data
```

Section 2: Append Panel Data

```
#Append g_2012 and g_2013
append_data = bind_rows(g_2012,g_2013)

#View the dimensions to check the append
dim(append_data)
```

```
## [1] 688   7
```

```
#Save the panel data in a Stata file
export(append_data, file = "panel_data.dta") #the appended 2012 and 2013 data frame
```

Section 3: Merge Data Sets

```
#Import the v5_guatemala_clean.dta file
v5_guatemala_cleaned = import("v5_guatemala_cleaned.dta")

#Look at the structure of the data
dim(v5_guatemala_cleaned)
```

```
## [1] 5070   40
```

```
str(v5_guatemala_cleaned)
```

```
## 'data.frame':    5070 obs. of  40 variables:
##  $ department           : chr  "alta verapaz" "alta verapaz" "alta verapaz" "alta verapaz" ...
##   ..- attr(*, "label")= chr "department"
##   ..- attr(*, "format.stata")= chr "%14s"
##  $ municipality         : chr  "cahabon" "cahabon" "cahabon" "cahabon" ...
##   ..- attr(*, "label")= chr "municipality"
##   ..- attr(*, "format.stata")= chr "%27s"
##  $ prev_mayor_ran_lost  : chr  "" "" "" "" ...
##   ..- attr(*, "label")= chr "prev_mayor_ran_lost"
##   ..- attr(*, "format.stata")= chr "%9s"
##  $ year                 : num  2000 2002 2004 2007 2008 ...
##   ..- attr(*, "label")= chr "year"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ unique_concat        : chr  "altaverapazcahabon" "altaverapazcahabon" "altaverapazcahabon" "altave
##   ..- attr(*, "label")= chr "unique_concat"
##   ..- attr(*, "format.stata")= chr "%40s"
##  $ crpt_infrac          : num  NA NA 2 5 3 1 4 3 5 5 ...
##   ..- attr(*, "label")= chr "infractions subcomponent: count"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ bureaucrats_         : num  NA NA 0 0 0 0 0 NA 0 0 ...
##   ..- attr(*, "label")= chr "bureaucrats subcomponent: count"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ complaints_          : num  NA NA 0 0 0 0 0 NA 0 0 ...
##   ..- attr(*, "label")= chr "complaints subcomponents: count"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ infrac_count_        : num  NA NA 2 5 3 1 4 3 5 5 ...
##   ..- attr(*, "label")= chr "total infractions, sum of subcomp: count"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ infrac_amount_       : num  NA NA 54054 185313 10228 ...
##   ..- attr(*, "label")= chr "(nom.)total infractions, sum of subcomp: amt"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ gov_transfers_       : num  NA NA NA NA 13807780 ...
##   ..- attr(*, "label")= chr "nom. sum of boost transfers, 2008-2015"
```

```
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ m_Gini_              : num  NA 24.4 NA NA NA ...
##   ..- attr(*, "label")= chr "Gini index in 2002 and 2011"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ m_extremepoverty_    : num  NA 61.1 NA NA NA ...
##   ..- attr(*, "label")= chr "% in extreme pov. in 2002 and 2011"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ m_totalpoverty_      : num  NA 93.8 NA NA NA ...
##   ..- attr(*, "label")= chr "% in tot. poverty in 2002 and 2011"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ pop_                 : num  NA 42797 46539 52153 54024 ...
##   ..- attr(*, "label")= chr "population imput. 1"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ Deflator             : num  NA 106 118 140 153 ...
##   ..- attr(*, "label")= chr "GDP deflator until 2017"
##   ..- attr(*, "format.stata")= chr "%10.0g"
##  $ m_extremepoverty_avg : num  NA 61.1 61.1 61.1 61.1 ...
##   ..- attr(*, "label")= chr "extreme poverty continuous var 2002-15"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ m_Gini_avg           : num  NA 24.4 24.4 24.4 24.4 ...
##   ..- attr(*, "label")= chr "Gini index continuous var. 2002-17"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ m_totalpoverty_avg   : num  NA 93.8 93.8 93.8 93.8 ...
##   ..- attr(*, "label")= chr "Gini index continuous var. 2002-17"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ pop_2_               : num  NA 42797 42797 54024 54024 ...
##   ..- attr(*, "label")= chr "population imput. 2"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ r_infrac_amount_     : num  NA NA 45828 132221 6668 ...
##   ..- attr(*, "label")= chr "real total infraction amount sum"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ r_gov_transfers_     : num  NA NA NA NA 9e+06 ...
##   ..- attr(*, "label")= chr "real boost govt transfres amt"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ r_crpt_amount_       : num  NA NA 45828 132221 6668 ...
##   ..- attr(*, "label")= chr "real infractions subcomp. amt "
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ r_complaints_amount_ : num  NA NA 0 0 0 0 NA NA NA NA ...
##   ..- attr(*, "label")= chr "real complaints subcomp. amt"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ r_bureaucrats_amount_: num  NA NA 0 0 0 0 NA NA NA NA ...
##   ..- attr(*, "label")= chr "real bureaucrats subcomp. amt"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ latitude             : num  15.6 15.6 15.6 15.6 15.6 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ longitude            : num  -89.8 -89.8 -89.8 -89.8 -89.8 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ term                 : num  1 1 2 2 3 3 3 3 4 4 ...
##   ..- attr(*, "label")= chr "term 1-5"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ female               : num  0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ mayorparty_votes_    : num  1937 1937 2969 2969 4444 ...
##   ..- attr(*, "label")= chr "votes won by mayor's party"
```

```
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ secparty_votes_      : num  1641 1641 1549 1549 3005 ...
##   ..- attr(*, "label")= chr "votes won by second place party"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ reelect_             : num  NA NA NA NA 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "party reelected"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ absdiff_             : num  296 296 1420 1420 1439 ...
##   ..- attr(*, "label")= chr "abs diff between first and second place party"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ percdiff_            : num  0.0488 0.0488 0.1687 0.1687 0.121 ...
##   ..- attr(*, "label")= chr "absdiff_ as share of valid_votes"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ reelect_ppl_         : num  NA NA 0 0 0 0 0 0 1 1 ...
##   ..- attr(*, "label")= chr "mayor reelected"
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ valid_votes          : num  6062 6062 8419 8419 11895 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ mayorparty_          : chr  "dia-urng" "dia-urng" "ppmrpsn" "ppmrpsn" ...
##   ..- attr(*, "label")= chr "winning mayor's party"
##   ..- attr(*, "format.stata")= chr "%25s"
##  $ secparty_            : chr  "frg" "frg" "urng" "urng" ...
##   ..- attr(*, "label")= chr "second place party"
##   ..- attr(*, "format.stata")= chr "%17s"
##  $ mayorname_           : chr  "" "" "ramiro iram garcia maaz" "ramiro iram garcia maaz" ...
##   ..- attr(*, "label")= chr "winning mayor name"
##   ..- attr(*, "format.stata")= chr "%50s"
##  $ d_prev_mayor_ran_lost: num  NA NA NA NA 0 0 0 0 NA NA ...
##   ..- attr(*, "label")= chr "dummy if prev mayor ran but lost"
##   ..- attr(*, "format.stata")= chr "%9.0g"
```

Rename `municipio` and `departamento` columns in `append_data` to ensure they merge as the same columns in the `v5_guatemala_clean` data:

```
#Rename the variables using the dplyr package
append_data = append_data %>%
  rename(department = departamento,
         municipality = municipio)

#Check to make sure the changes went through
colnames(append_data)
```

```
## [1] "department"              "municipality"
## [3] "tasa_bruta_de_natalidad" "tasa_de_criminalidad"
## [5] "tasa_de_homicidios"      "violencia_intrafamiliar"
## [7] "year"
```

```
#Adjust the labels with the labelled package
var_label(append_data) <- list(`department` = "Department",
                           `municipality` = "Municipality",
                           `tasa_bruta_de_natalidad` = "Birth Rate",
                           `tasa_de_criminalidad` = "Crime Rate",
                           `tasa_de_homicidios` = "Homicide Rate",
```

```
                          `violencia_intrafamiliar` = "Domestic Violence",
                          `year` = "Year")
```

Merge the data:

```
#merge append_data and v5_guatemala_clean data
v6_Guatemala = left_join(x = v5_guatemala_cleaned, y = append_data)
```

```
## Joining, by = c("department", "municipality", "year")
```

```
#check the merge
dim(v6_Guatemala)
```

```
## [1] 5070   44
```

```
colnames(v6_Guatemala)
```

```
##  [1] "department"              "municipality"
##  [3] "prev_mayor_ran_lost"     "year"
##  [5] "unique_concat"           "crpt_infrac"
##  [7] "bureaucrats_"            "complaints_"
##  [9] "infrac_count_"           "infrac_amount_"
## [11] "gov_transfers_"          "m_Gini_"
## [13] "m_extremepoverty_"       "m_totalpoverty_"
## [15] "pop_"                    "Deflator"
## [17] "m_extremepoverty_avg"    "m_Gini_avg"
## [19] "m_totalpoverty_avg"      "pop_2_"
## [21] "r_infrac_amount_"        "r_gov_transfers_"
## [23] "r_crpt_amount_"          "r_complaints_amount_"
## [25] "r_bureaucrats_amount_"   "latitude"
## [27] "longitude"               "term"
## [29] "female"                  "mayorparty_votes_"
## [31] "secparty_votes_"         "reelect_"
## [33] "absdiff_"                "percdiff_"
## [35] "reelect_ppl_"            "valid_votes"
## [37] "mayorparty_"             "secparty_"
## [39] "mayorname_"              "d_prev_mayor_ran_lost"
## [41] "tasa_bruta_de_natalidad" "tasa_de_criminalidad"
## [43] "tasa_de_homicidios"      "violencia_intrafamiliar"
```

```
#Save the new data set as a Stata file
export(v6_Guatemala, file = "v6_Guatemala.dta")
```

Section 4: Prepare Specifications

Specification context

Domestic violence is the outcome that I'm interested in analyzing. I consider domestic violence the "x" variable in my future analyses. I consider corruption, mayor gender, crime rate, and homicide rate as potential "y" variables that may explain variation in "x" values (i.e. domestic violence). Note: This data

set only includes domestic violence, homicide rates, and crime rates for years 2012 and 2013. I hope to add more years of these municipal-level data in the future.

For quick reference: `violencia_intrafamiliar` is the variable to account for domestic violence. `infrac_amount_` is the variable to account for corruption. `tasa_de_homicidios` is the variable to account for homicide rates. `tasa_de_criminalidad` is the variable to account for crime rates. `female` is a dummy variable to account for mayor gender.

```
#Specification 1: corruption and domestic violence
specification1 = v6_Guatemala %>% drop_na("violencia_intrafamiliar",
                                          "infrac_amount_")
dim(specification1)
```

```
## [1] 632  44
```

```
#Specification 2: mayor gender and domestic violence
specification2 = v6_Guatemala %>% drop_na("violencia_intrafamiliar",
                                          "female")
dim(specification2)
```

```
## [1] 630  44
```

```
#Specification 3: corruption, mayor gender, and domestic violence
specification3 = v6_Guatemala %>% drop_na("infrac_amount_",
                                          "female", "violencia_intrafamiliar")
dim(specification3)
```

```
## [1] 630  44
```

```
#Specification 4: crime rate, homicide rate, and domestic violence
specification4 = v6_Guatemala %>% drop_na("tasa_de_criminalidad",
                                          "tasa_de_homicidios",
                                          "violencia_intrafamiliar")
dim(specification4)
```

```
## [1] 502  44
```

Section 5: Initialize to Git

To see the R Markdown version of this document, clink on this link to access its GitHub repository - "GuatemalaProject."