Final Project

2022-11-09

Contents

1 Data wrangling

1. Setting path containing Peruvian dataframes:

```
path <- "/Users/khristelzavaleta/Desktop/Uchicago/Q4/Data and programming II/Homework/final-project-dieg
setwd("/Users/khristelzavaleta/Desktop/Uchicago/Q4/Data and programming II/Homework/final-project-diego
```

1.1 World Bank femicides data

2. Automatically retrieved dataset

```
femicides <- WDI(
  indicator = "VC.IHR.PSRC.FE.P5", country = c("MX", "PE"),
  start = 2011, end = 2020
)


write.csv(femicides, paste0(path,"/Data/final_dataframes/femicides_mexicoperu.csv"))
```

1.2 Peru data wrangling

3. Loading national poll:

```
enares_2019 <- read_dta(paste0(path, "/Data/Peru/14_v_c2cap400.dta"))
enares_2019_2 <- read_dta(paste0(path, "/Data/Peru/12_v_c2cap200.dta"),
  encoding = "latin1"
)

enares_2015 <- read.dbf(paste0(path, "/Data/Peru/08_CRS02_CAP400.dbf"))

enares_2015_2 <- read.dbf(paste0(path, "/Data/Peru/06_CRS02_CAP200.dbf"))

enares_2013 <- read.dbf(paste0(path, "/Data/Peru/11_CRS02_CAP400.dbf")) %>%
  rename("CCDD" = "C2CCDD")
enares_2013_2 <- read.dbf(paste0(path, "/Data/Peru/09_CRS02_CAP200.dbf"))
```

4. Function to clean data and select variables of interest

```
dfunction <- function(df, postpone_goals, obey, sexrelations, cheating, year) {
  names(df)[length(df)] <- "factor"

  regions <- read_xlsx(paste0(path, "/Data/Peru/geodir-ubigeo-inei.xlsx"))
  regions$region <- substr(regions$Ubigeo, 1, 2)

  regions_2 <- regions %>%
    group_by(region) %>%
    count(Departamento)

  df_1 <- df %>%
    mutate(
      w_postpone_goals = ifelse(postpone_goals == 1 | postpone_goals == 2, 0, 1),
      w_obey = ifelse(obey == 1 | obey == 2, 0, 1),
      w_willing_sex = ifelse(sexrelations == 1 | sexrelations == 2, 0, 1),
      punish_cheating = ifelse(cheating == 1 | cheating == 2, 0, 1),
      year_poll = year
    )

  df_2 <- merge(df_1, regions_2[, c("region", "Departamento")],
      by.x = c("CCDD"),
      by.y = c("region")
    )

  return(df_2)
}

# Function to add fix effects

dfunction_2 <- function(df) {
  setnames(df, replace(names(df), c(
    length(df), length(df) - 1, length(df) - 2,
    length(df) - 3, length(df) - 4
  ), c("sex", "marital_status", "employed", "education_level", "years_old")))
}
```

5. Applying the function with the specific columns codes for each year poll

5.1. Year 2019

```
peru_2019 <- dfunction(
  enares_2019, enares_2019$C2P401_10, enares_2019$C2P401_9,
  enares_2019$C2P401_7, enares_2019$C2P401_5, 2019
) %>% select(
  ID, HOGAR_ID, PERSONA_ID, w_postpone_goals, w_obey, w_willing_sex, punish_cheating,
  year_poll, factor, Departamento
)

# Merging data related to marital status, studies, years old, employed
peru_2019 <- merge(peru_2019, enares_2019_2[, c(
  "ID", "HOGAR_ID", "PERSONA_ID",
  "C1P208_A", "C1P210", "C1P211", "C1P212", "C1P207"
)],
by = c("ID", "HOGAR_ID", "PERSONA_ID"), all.x = TRUE
)

peru_2019 <- dfunction_2(peru_2019) %>%
  add_column(., CONGLOMERA = NA, .before = "ID") #others df have 4 key columns
```

5.2. Year 2015

```
#Year 2015

peru_2015 <- dfunction(
  enares_2015, enares_2015$C2P403_1, enares_2015$C2P406_1,
  enares_2015$C2P407_3, enares_2015$C2P411_2, 2015
) %>%
  select(
    CONGLOMERA, NSELV, HOGARN, PERSONA_ID, w_postpone_goals, w_obey,
    w_willing_sex, punish_cheating, year_poll,
    factor, Departamento
  )

peru_2015 <- merge(peru_2015, enares_2015_2[, c(
  "CONGLOMERA", "NSELV", "HOGARN",
  "PERSONA_ID", "C2P208_A", "C2P210", "C2P211", "C2P212", "C2P207"
)],
by = c("CONGLOMERA", "NSELV", "HOGARN", "PERSONA_ID")
)

peru_2015 <- dfunction_2(peru_2015)
peru_2015$CONGLOMERA <- as.character(peru_2015$CONGLOMERA)
```

5.3. Year 2013

```
peru_2013 <- dfunction(
  enares_2013, enares_2013$C2P4031, enares_2013$C2P4061,
  enares_2013$C2P4073, enares_2013$C2P4112, 2013
) %>%
  select(
    C2CONGLOME, C2NSELV, C2HOGARN, C2P201, w_postpone_goals, w_obey,
    w_willing_sex, punish_cheating, year_poll,
    factor, Departamento
  )


peru_2013 <- merge(peru_2013, enares_2013_2[, c(
  "C2CONGLOME", "C2NSELV", "C2HOGARN", "C2P201", "C2P208ANIO",
  "C2P210", "C2P211", "C2P212", "C2P207"
)],
by = c("C2CONGLOME", "C2NSELV", "C2HOGARN", "C2P201")
)

peru_2013 <- dfunction_2(peru_2013)
peru_2013$C2CONGLOME <- as.character(peru_2013$C2CONGLOME)
```

6. Binding data from 2013, 2015 and 2019

```
peru_data <- as.data.frame(mapply(c, peru_2013,peru_2015, peru_2019))

peru_data$employed <- replace(peru_data$employed, peru_data$employed == 2, 0)
peru_data$education_level <- as.numeric(peru_data$education_level) - 1
peru_data[, 5:9] <- lapply(peru_data[, 5:9], as.numeric)
peru_data <- peru_data[, c(9,11,1:8, 12:16, 10)] #Reorder column by position
```

7. Creating Peruvian Index of social tolerance to violence (per year)

```
peru_data_2 <- peru_data %>%
  group_by(Departamento, year_poll, factor) %>%
  summarise(
    punish_cheating = wtd.mean(punish_cheating, as.numeric(factor)),
    w_postpone_goals = wtd.mean(w_postpone_goals, as.numeric(factor)),
    w_obey = wtd.mean(w_obey, as.numeric(factor)),
    w_willing_sex = wtd.mean(w_willing_sex, as.numeric(factor))
  ) %>%
  group_by(Departamento, year_poll) %>%
  summarise(
    punish_cheating = mean(punish_cheating), w_postpone_goals = mean(w_postpone_goals),
    w_obey = mean(w_obey), w_willing_sex = mean(w_willing_sex)
  ) %>%
  group_by(Departamento, year_poll) %>%
  mutate(Index = sum(punish_cheating, w_postpone_goals, w_obey, w_willing_sex) / 4) %>%
  mutate(across(where(is.numeric), round, 3))
```

8. Merging data with the dependent variables

```
dependent_variables <- c(
  "peru_violencia_sexual", "peru_violencia_psicologica",
  "peru_violencia_fisica"
)
rep_str <- c(
  "Áncash" = "Ancash", "Apurímac" = "Apurimac", "Huánuco" = "Huanuco",
  "Junín" = "Junin", "San Martín" = "San Martin"
)


for (i in dependent_variables) {
  assign(i, read_xlsx(paste0(path,"/Data/Peru/",i, ".xlsx")) %>%
       mutate_at(c(2:14), as.numeric) %>%
  mutate(across(where(is.numeric), round, 2)) %>%
  select("Ámbito geográfico", "2013", "2015", "2019") %>%
  pivot_longer("2013":"2019", names_to = "year", values_to = i) %>%
  mutate(`Ámbito geográfico` = str_replace_all(`Ámbito geográfico`, rep_str)))
}

data = c("peru_data_2", "peru_data")

peru_data_2 <- peru_data_2 %>%
  merge(peru_violencia_sexual,
    by.x = c("Departamento", "year_poll"),
    by.y = c("Ámbito geográfico", "year"), all.x = TRUE) %>%
  merge(peru_violencia_psicologica,
    by.x = c("Departamento", "year_poll"),
    by.y = c("Ámbito geográfico", "year"), all.x = TRUE) %>%
  merge(peru_violencia_fisica,
    by.x = c("Departamento", "year_poll"),
    by.y = c("Ámbito geográfico", "year"), all.x = TRUE)

peru_data_long <- peru_data %>%
merge(peru_violencia_sexual,
    by.x = c("Departamento", "year_poll"),
    by.y = c("Ámbito geográfico", "year"), all.x = TRUE) %>%
  merge(peru_violencia_psicologica,
    by.x = c("Departamento", "year_poll"),
    by.y = c("Ámbito geográfico", "year"), all.x = TRUE) %>%
  merge(peru_violencia_fisica,
    by.x = c("Departamento", "year_poll"),
    by.y = c("Ámbito geográfico", "year"), all.x = TRUE)

peru_data_long <- peru_data_long %>%
  mutate(Index = (w_obey * as.numeric(factor) + w_postpone_goals * as.numeric(factor) +
    w_willing_sex * as.numeric(factor) +
    punish_cheating * as.numeric(factor)) / (as.numeric(factor) * 4))
```

9. Save the peruvian dataframe as a csv file

```
write.csv(peru_data_2, paste0(path, "/Data/final_dataframes/peru_data.csv"),
  row.names = FALSE
)
write.csv(peru_data_long, paste0(path, "/Data/final_dataframes/peru_data_long.csv"),
  row.names = FALSE
)
```

1.3 Mexico data wrangling

1. Reading Data

```
mexico_data <- read_csv(paste0(path, "/Data/final_dataframes/mexico_data.csv"))

rep_str_mexico <- c(
  "Estado de mexico" = "México", "Mexico" = "México", "Baja california" = "Baja California",
  "Baja california sur" = "Baja California Sur", "Ciudad de mexico" = "Ciudad de México",
  "Coahuila de zaragoza" = "Coahuila de Zaragoza", "Michoacan de ocampo" = "Michoacán de Ocampo",
  "Nuevo leon" = "Nuevo León", "Queretaro" = "Querétaro", "Quintana roo" = "Quintana Roo",
  "San luis potosi" = "San Luis Potosí",
  "Veracruz de ignacio de la llave" = "Veracruz de Ignacio de la Llave", "Yucatan" = "Yucatán"
)

mexico_data <- mexico_data %>%
  drop_na(w_willing_sex, w_house_chores, w_chooseto_work_study, w_conflict_jelousy) %>%
  mutate(state = str_replace_all(state, rep_str_mexico))

mexico_data$w_conflict_jelousy[mexico_data$year_poll == 2021] <-
  ifelse(mexico_data$w_conflict_jelousy[mexico_data$year_poll == 2021] == 1, 0, 1)

mexico_data$w_house_chores[mexico_data$year_poll == 2021] <-
  ifelse(mexico_data$w_house_chores[mexico_data$year_poll == 2021] == 1, 0, 1)

mexico_data$w_chooseto_work_study[mexico_data$year_poll == 2021] <-
  ifelse(mexico_data$w_chooseto_work_study[mexico_data$year_poll == 2021] == 1, 0, 1)
```

```r
mexico_data$w_chooseto_work_study <- ifelse(mexico_data$w_chooseto_work_study == 0, 1, 0)


mexico_data <- mexico_data %>%
  mutate(state = str_replace_all(state, c("Baja California sur" = "Baja California Sur")))
```

2. Creating Mexican Index

```r
mexico_data_short <- mexico_data %>%
  group_by(state, year_poll, FAC_MUJ) %>%
  summarise(
    w_willing_sex = wtd.mean(w_willing_sex, FAC_MUJ),
    w_house_chores = wtd.mean(w_house_chores, FAC_MUJ),
    w_chooseto_work_study = wtd.mean(w_chooseto_work_study, FAC_MUJ),
    w_conflict_jelousy = wtd.mean(w_conflict_jelousy, FAC_MUJ)
  ) %>%
  group_by(state, year_poll) %>%
  summarise(
    w_willing_sex = mean(w_willing_sex), w_house_chores = mean(w_house_chores),
    w_chooseto_work_study = mean(w_chooseto_work_study),
    w_conflict_jelousy = mean(w_conflict_jelousy)
  ) %>%
  group_by(state, year_poll) %>%
  mutate(Index = sum(
    w_willing_sex, w_house_chores, w_chooseto_work_study,
    w_conflict_jelousy
  ) / 4) %>%
  mutate(across(where(is.numeric), round, 3))


mex_summ <- mexico_data %>%
  group_by(state, year_poll) %>%
  summarise(
    mexico_violencia_psicologica = mean(mexico_violencia_psicologica),
    mexico_violencia_fisica = mean(mexico_violencia_fisica),
    mexico_violencia_sexual = mean(mexico_violencia_sexual)
  )

mexico_data_short <- mexico_data_short %>%
  merge(mex_summ, by = c("state", "year_poll"))


mexico_data <- mexico_data %>%
  mutate(Index = (w_willing_sex + w_house_chores +
    w_chooseto_work_study + w_conflict_jelousy) / 4)
```

3. Save the mexican dataframe as a csv file

```r
write.csv(mexico_data, paste0(path, "/Data/final_dataframes/mexico_data_long.csv"),
  row.names = FALSE
)
write.csv(mexico_data_short, paste0(path, "/Data/final_dataframes/mexico_data_short.csv"),
  row.names = FALSE
)
```

2 Plotting

1. Shape files to be merge with the data

- Peru

```
peru_shapefile <- st_read(paste0(
  path,
  "/Data/Peru/Peru_shapefile/per_admbnda_adm1_ign_20200714.shp"
))
```

```
## Reading layer `per_admbnda_adm1_ign_20200714' from data source
##   `/Users/khristelzavaleta/Desktop/Uchicago/Q4/Data and programming II/Homework/final-project-diego_
##   using driver `ESRI Shapefile'
## Simple feature collection with 25 features and 13 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -81.32823 ymin: -18.35093 xmax: -68.65228 ymax: -0.03860597
## Geodetic CRS:  WGS 84
```

```
peru_shapefile <- st_transform(peru_shapefile, 4326)
```

```
peru_data_sf <- peru_data_2 %>%
  merge(peru_shapefile[, c("ADM1_ES", "geometry")],
    by.x = c("Departamento"),
    by.y = c("ADM1_ES"), all.y = TRUE
  )

peru_data_sf <- st_sf(peru_data_sf)
```

- Mexico

```
mexico_shapefile <- st_read(paste0(path, "/Data/Mexico/mexico_shapefile/01_32_ent.shp"))
```

```
## Reading layer `01_32_ent' from data source
##   `/Users/khristelzavaleta/Desktop/Uchicago/Q4/Data and programming II/Homework/final-project-diego_
##   using driver `ESRI Shapefile'
## Simple feature collection with 32 features and 3 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 911292 ymin: 319149.1 xmax: 4082997 ymax: 2349615
## Projected CRS: MEXICO_ITRF_2008_LCC
```

```
mexico_shapefile <- st_transform(mexico_shapefile, 4326)
```

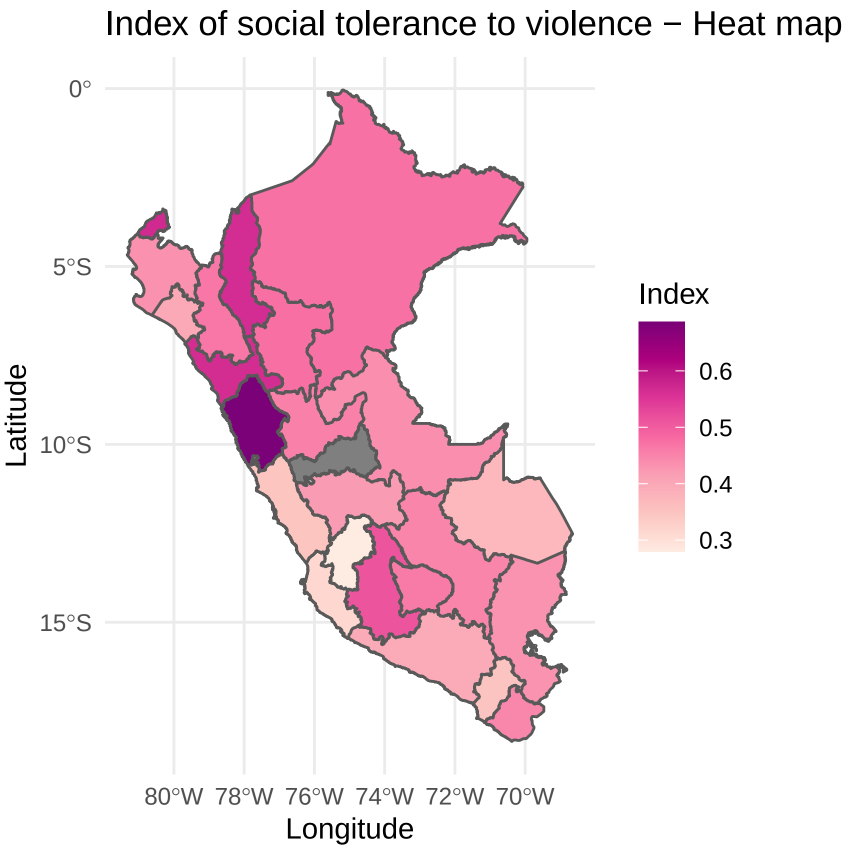
```
mexico_data_short_sf <- mexico_data_short %>%
  merge(mexico_shapefile[, c("NOMGEO", "geometry")],
    by.x = c("state"),
    by.y = c("NOMGEO"), all.x = TRUE
  )

mexico_data_short_sf <- st_sf(mexico_data_short_sf)
```

2.1 R plots

2.1.1 Maps

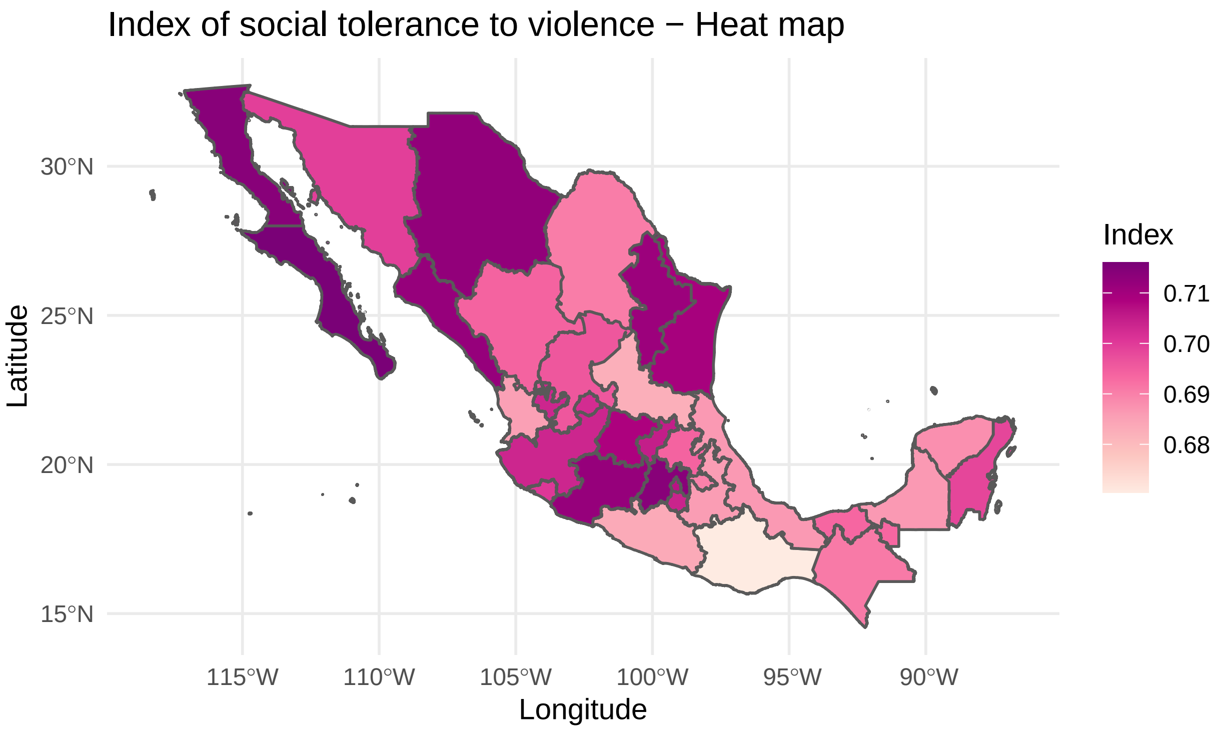
- Heat Map - Peru

```
plot_map_peru <- peru_data_sf %>%
  group_by(Departamento) %>%
  summarise(Index = mean(Index)) %>%
  ggplot() +
  geom_sf(aes(fill = Index)) +
  ggtitle("Index of social tolerance to violence - Heat map ") +
  labs(y = "Latitude", x = "Longitude") +
  scale_fill_distiller(palette = "RdPu", direction = 1) +
  theme_minimal()

ggsave(filename = paste0(path, "/images/plot_map_peru.png"), plot = plot_map_peru)

plot_map_peru
```


Index of social tolerance to violence − Heat map
0°
5°S
10°S
15°S
Latitude
80°W 78°W 76°W 74°W 72°W 70°W
Longitude
Index
0.6
0.5
0.4
0.3

- Heat Map - Mexico

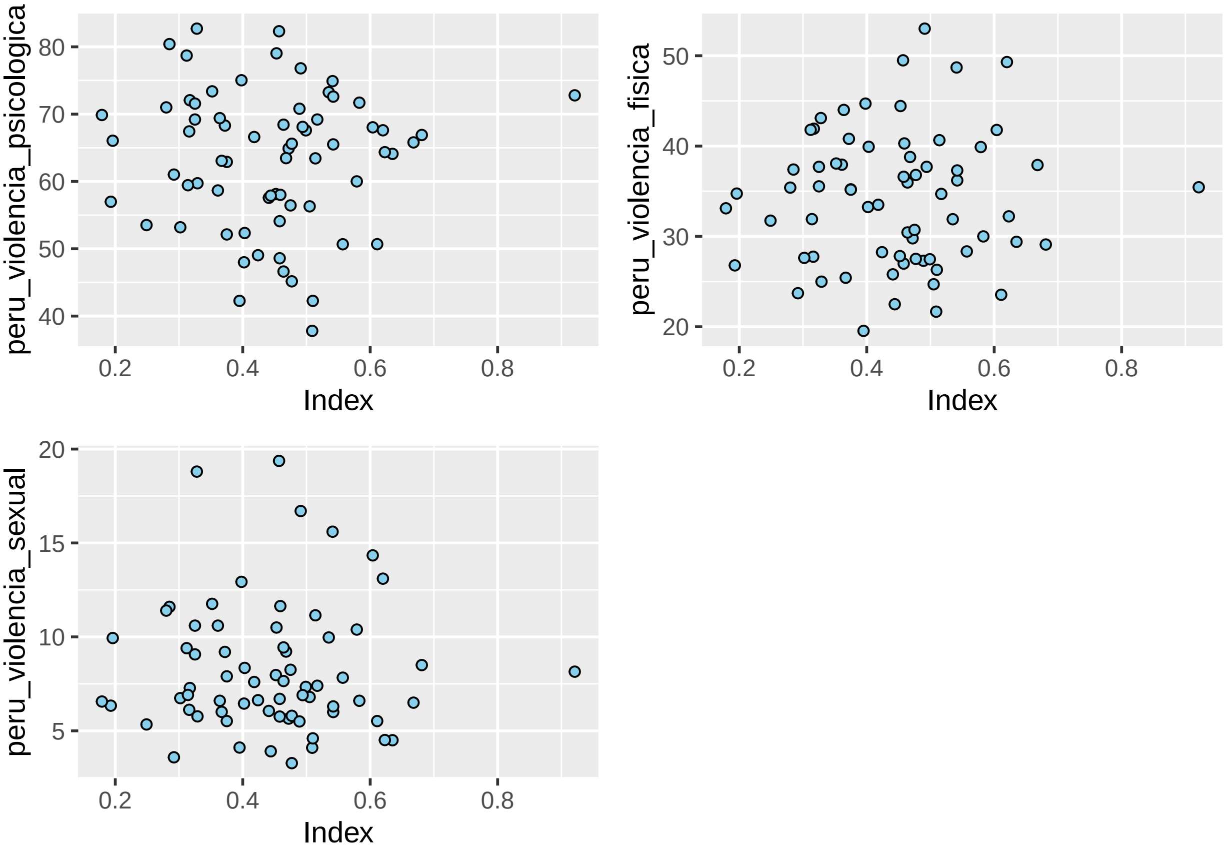
```
plot_map_mexico <- mexico_data_short_sf %>%
  group_by(state) %>%
  summarise(Index = mean(Index)) %>%
  ggplot() +
  geom_sf(aes(fill = Index)) +
  ggtitle("Index of social tolerance to violence - Heat map ") +
  labs(y = "Latitude", x = "Longitude") +
  scale_fill_distiller(palette = "RdPu", direction = 1) +
  theme_minimal()

ggsave(filename = paste0(path, "/images/plot_map_mexico.png"), plot = plot_map_mexico)

plot_map_mexico
```



2.1.2 Scatter plot

- Exploratory analysis : Index vs dependent variables - Peru

```
violencia_psicologica <- ggplot(data = peru_data_2, aes(
  x = Index,
  y = peru_violencia_psicologica
)) +
  geom_point(fill = "skyblue", shape = 21)

violencia_fisica <- ggplot(data = peru_data_2, aes(
  x = Index,
  y = peru_violencia_fisica
)) +
  geom_point(fill = "skyblue", shape = 21)

violencia_sexual <- ggplot(data = peru_data_2, aes(
  x = Index,
  y = peru_violencia_sexual
)) +
  geom_point(fill = "skyblue", shape = 21)

plot_scatter_peru <- grid.arrange(violencia_psicologica, violencia_fisica,
  violencia_sexual,
  ncol = 2
)
```

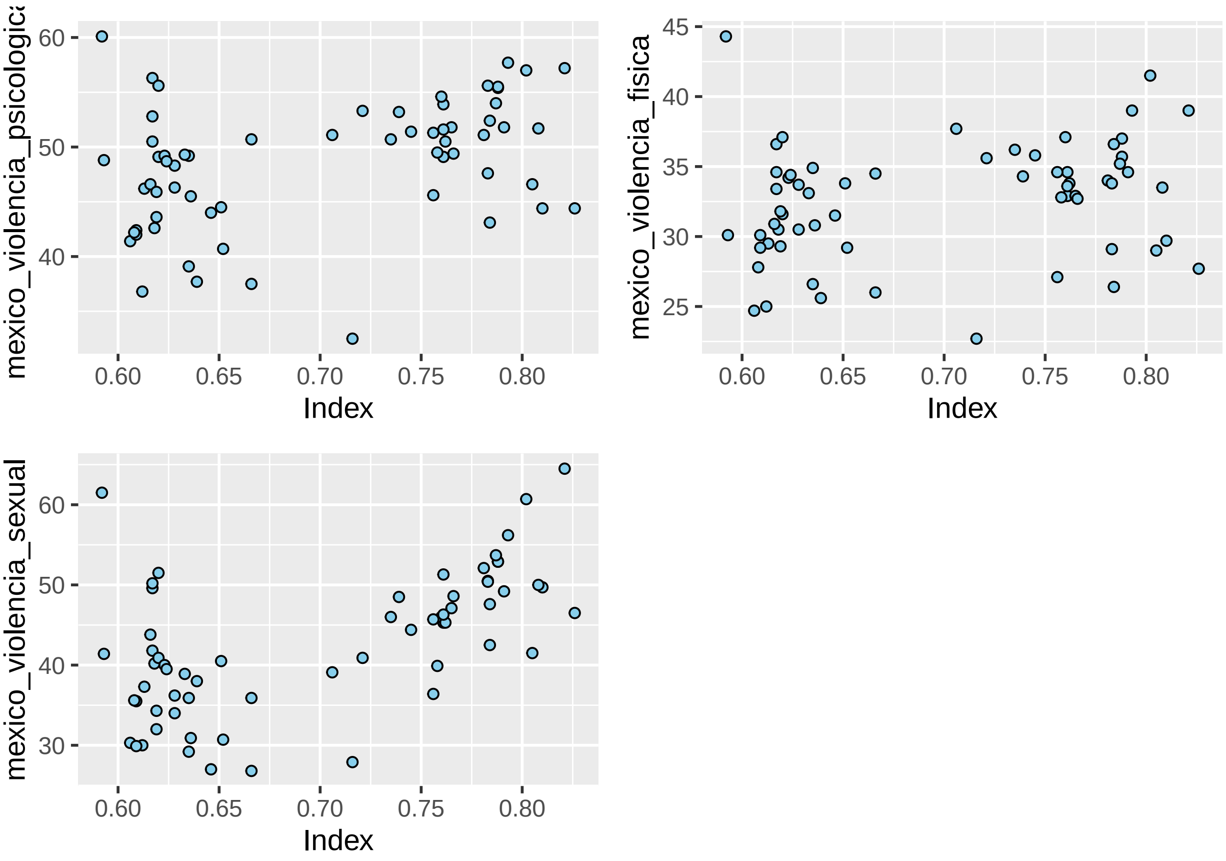


```
ggsave(
  filename = paste0(path, "/images/plot_scatter_peru.png"),
  plot = plot_scatter_peru
)
```

- Exploratory analysis : Index vs dependent variables - Mexico

```
violencia_psicologica_mx <- ggplot(data = mexico_data_short, aes(
  x = Index,
  y = mexico_violencia_psicologica
)) +
  geom_point(fill = "skyblue", shape = 21)
```

```
violencia_fisica_mx <- ggplot(data = mexico_data_short, aes(
  x = Index,
  y = mexico_violencia_fisica
)) +
  geom_point(fill = "skyblue", shape = 21)

violencia_sexual_mx <- ggplot(data = mexico_data_short, aes(
  x = Index,
  y = mexico_violencia_sexual
)) +
  geom_point(fill = "skyblue", shape = 21)

plot_scatter_mexico <- grid.arrange(violencia_psicologica_mx, violencia_fisica_mx,
  violencia_sexual_mx,
  ncol = 2
)
```

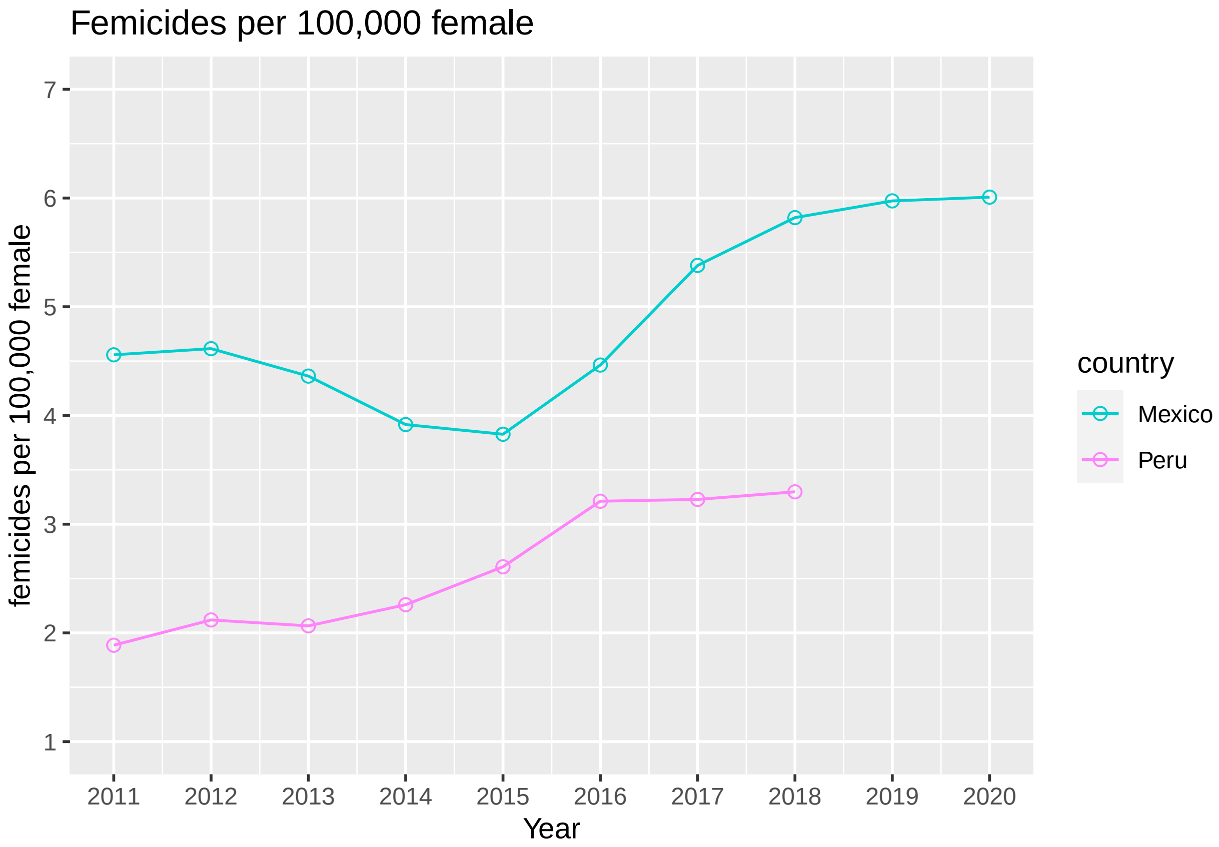


```
ggsave(
  filename = paste0(path, "/images/plot_scatter_mexico.png"),
  plot = plot_scatter_mexico
)
```

2.1.3 Lines plot

- Femicides Mexico and Peru

```
plot_femicides <- ggplot(femicides, aes(
  x = year, y = VC.IHR.PSRC.FE.P5,
  group = country, color = country
)) +
  geom_line() +
  scale_color_manual(values = c("cyan3", "#FF83FA")) +
  scale_y_continuous(breaks = seq(1, 7, by = 1), limits = c(1, 7)) +
  scale_x_continuous(breaks = seq(2011, 2020, by = 1), limits = c(2011, 2020)) +
  ggtitle("Femicides per 100,000 female") +
  # theme_ipsum() +
  labs(y = "femicides per 100,000 female", x = "Year") +
  geom_point(size = 2, shape = 21)

ggsave(filename = paste0(path, "/images/plot_femicides.png"), plot = plot_femicides)

plot_femicides
```



2.1.4 Animated plot

- Animation Peru - Progression analysis : Index vs sexual violence (colored by state)

```r
per_dep_var <- c(
  "peru_violencia_sexual", "peru_violencia_fisica",
  "peru_violencia_psicologica"
)


for (i in per_dep_var) {
  assign(
    i, ggplot(data = peru_data_2) +
      geom_point(aes(x = Index, y = peru_data_2[, c(i)], fill = Departamento, size = Index),
        shape = 21,
        alpha = 0.5
      ) +
      scale_size(range = c(1, 10)) +
      scale_fill_viridis(discrete = TRUE, guide = "none", option = "A")
    #+ theme_ipsum()
  )
}


gif_peru_1 <- peru_violencia_sexual +
  labs(x = "Index", y = "Sexual violence", title = "Index vs Sexual violence") +
  transition_time(as.integer(year_poll)) +
  labs(title = "Peru - Index vs Sexual violence progression by year: {frame_time}")

gif_peru_2 <- peru_violencia_fisica +
  labs(x = "Index", y = "Physical violence", title = "Index vs Physical violence") +
  transition_time(as.integer(year_poll)) +
  labs(title = "Peru - Index vs Physical violence progression by year: {frame_time}")

gif_peru_3 <- peru_violencia_psicologica +
  labs(x = "Index", y = "Psychological violence", title = "Index vs Psychological violence") +
  transition_time(as.integer(year_poll)) +
  labs(title = "Peru - Index vs Psychological violence progression by year: {frame_time}")

anim_save(paste0(path, "/images/gif_peru_1.gif"),
  animation = gif_peru_1,
  height = 400, width = 500
)
anim_save(paste0(path, "/images/gif_peru_2.gif"),
  animation = gif_peru_2,
  height = 400, width = 500
)
anim_save(paste0(path, "/images/gif_peru_3.gif"),
  animation = gif_peru_3,
  height = 400, width = 500
)
```

- Animation Mexico - Progression analysis : Index vs sexual violence (colored by state)

```r
require(ggplot2)
mex_dep_var <- c(
  "mexico_violencia_sexual", "mexico_violencia_fisica",
  "mexico_violencia_psicologica"
)


for (i in mex_dep_var) {
  assign(
    i, ggplot(data = mexico_data_short) +
      geom_point(aes(x = Index, y = mexico_data_short[, c(i)], fill = state, size = Index),
        shape = 21,
        alpha = 0.5
      ) +
      scale_size(range = c(1, 10)) +
      scale_fill_viridis(discrete = TRUE, guide = "none", option = "A")
    # + theme_ipsum()
  )
}


gif_mexico_1 <- mexico_violencia_sexual +
  labs(x = "Index", y = "Sexual violence", title = "Index vs Sexual violence") +
  transition_time(as.integer(year_poll)) +
  labs(title = "Mexico - Index vs Sexual violence progression by year: {frame_time}")

gif_mexico_2 <- mexico_violencia_fisica +
  labs(x = "Index", y = "Physical violence", title = "Index vs Physical violence") +
  transition_time(as.integer(year_poll)) +
  labs(title = "Mexico - Index vs Physical violence progression by year: {frame_time}")

gif_mexico_3 <- mexico_violencia_psicologica +
  labs(x = "Index", y = "Psychological violence", title = "Index vs Psychological violence") +
  transition_time(as.integer(year_poll)) +
  labs(title = "Mexico - Index vs Psychological violence progression by year: {frame_time}")

anim_save(paste0(path, "/images/gif_mexico_1.gif"),
  animation = gif_mexico_1,
  height = 400, width = 500
)
anim_save(paste0(path, "/images/gif_mexico_2.gif"),
  animation = gif_mexico_2,
  height = 400, width = 500
)
anim_save(paste0(path, "/images/gif_mexico_3.gif"),
  animation = gif_mexico_3,
  height = 400, width = 500
)
```

- Animation femicides

```
require(ggplot2)

gif_2 <- plot_femicides +
  transition_reveal(year) +
 labs(title = "Femicides per 100,000 female: {frame_along}")
```

```
anim_save(paste0(path,"/images/gif_2.gif"), animation = gif_2)

gif_2
```

- Image created for the home page of shiny

```
america <- ne_countries(continent = c('south america', "north america"), scale = "small",
                        returnclass = "sf")

peru <- ne_countries(country = "peru", scale = "small", returnclass = "sf")
mexico <- ne_countries(country = "mexico", scale = "small", returnclass = "sf")

plot_peru_mexico <- ggplot() +
  geom_sf(data = america[(america$sovereignt != "Canada" &
    america$sovereignt != "United States of America" &
      america$sovereignt != "Denmark"), ]) +
  geom_sf(data = peru, fill = "red") +
  geom_sf(data = mexico, fill = "red") +
  theme_bw() +
  theme(axis.text.x = element_blank(), axis.text.y = element_blank())

ggsave(plot = plot_peru_mexico, filename = paste0(path, "/www/plot_peru_mexico.jpg"))
```

```
## Saving 6.5 x 4.5 in image
```

2.2 Shiny code

Not in this document

3 Text processing

```
# load the policy documents from both Mexico and Peru

mimp_peru <- pdf_text(paste0(path, "/Data/Peru/MIMP-violencia-basada_en_genero.pdf"))

mpg_mexico <- pdf_text(paste0(path, "/Data/Mexico/Manual_Violencia_de_G_nero_en_Diversos_Contextos2.pdf
```

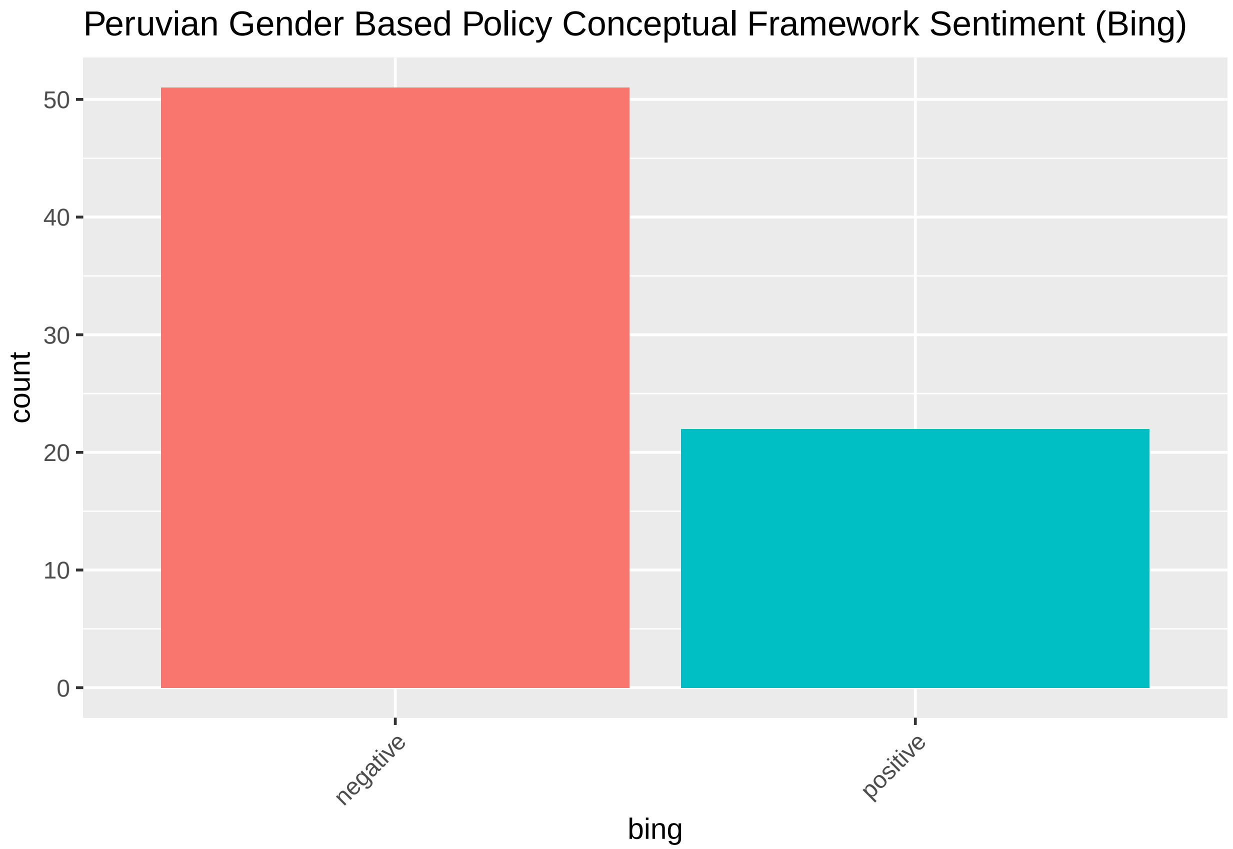
```
#Separating words, sentences and ngrams

mexico <- tibble(text = mpg_mexico)
word_tokens_mexico <- unnest_tokens(mexico, word_tokens, text, token = "words")
sentence_tokens_mexico <- unnest_tokens(mexico, sent_tokens, text, token = "sentences")
ngram_tokens_mexico <- unnest_tokens(mexico, ngram_tokens, text, token = "ngrams", n = 2)

peru <- tibble(text = mimp_peru)
word_tokens_peru <- unnest_tokens(peru, word_tokens, text, token = "words")
sentence_tokens_peru <- unnest_tokens(peru, sent_tokens, text, token = "sentences")
ngram_tokens_peru <- unnest_tokens(peru, ngram_tokens, text, token = "ngrams", n = 2)
```

```
# Adding Spanish words to stop words using the tm library
stop_words_spanish <- bind_rows(
  stop_words,
  data_frame(
    word = stopwords("spanish"),
    lexicon = "custom"
  )
)

peru_no_sw <- anti_join(word_tokens_peru, stop_words_spanish,
  by = c("word_tokens" = "word")
)

mexico_no_sw <- anti_join(word_tokens_mexico, stop_words_spanish,
  by = c("word_tokens" = "word")
)
```

```
#Analyzing sentiments
sentiment_nrc <-
  get_sentiments("nrc") %>%
  rename(nrc = sentiment)
sentiment_afinn <-
  get_sentiments("afinn") %>%
  rename(affin = value)
sentiment_bing <-
  get_sentiments("bing") %>%
  rename(bing = sentiment)
```

3.1 Plotting the Peruvian policy document

```
# Plotting the Peruvian policy document

peru_no_sw <- peru_no_sw %>%
  left_join(sentiment_nrc, by = c("word_tokens" = "word")) %>%
  left_join(sentiment_afinn, by = c("word_tokens" = "word")) %>%
  left_join(sentiment_bing, by = c("word_tokens" = "word"))

# Sentiment bing
ggplot(data = filter(peru_no_sw, !is.na(bing))) +
  geom_histogram(aes(bing, fill = bing), stat = "count") +
  scale_x_discrete(guide = guide_axis(angle = 45)) +
  labs(
    title =
      "Peruvian Gender Based Policy Conceptual Framework Sentiment (Bing)"
  ) +
  theme(legend.position = "none")
```

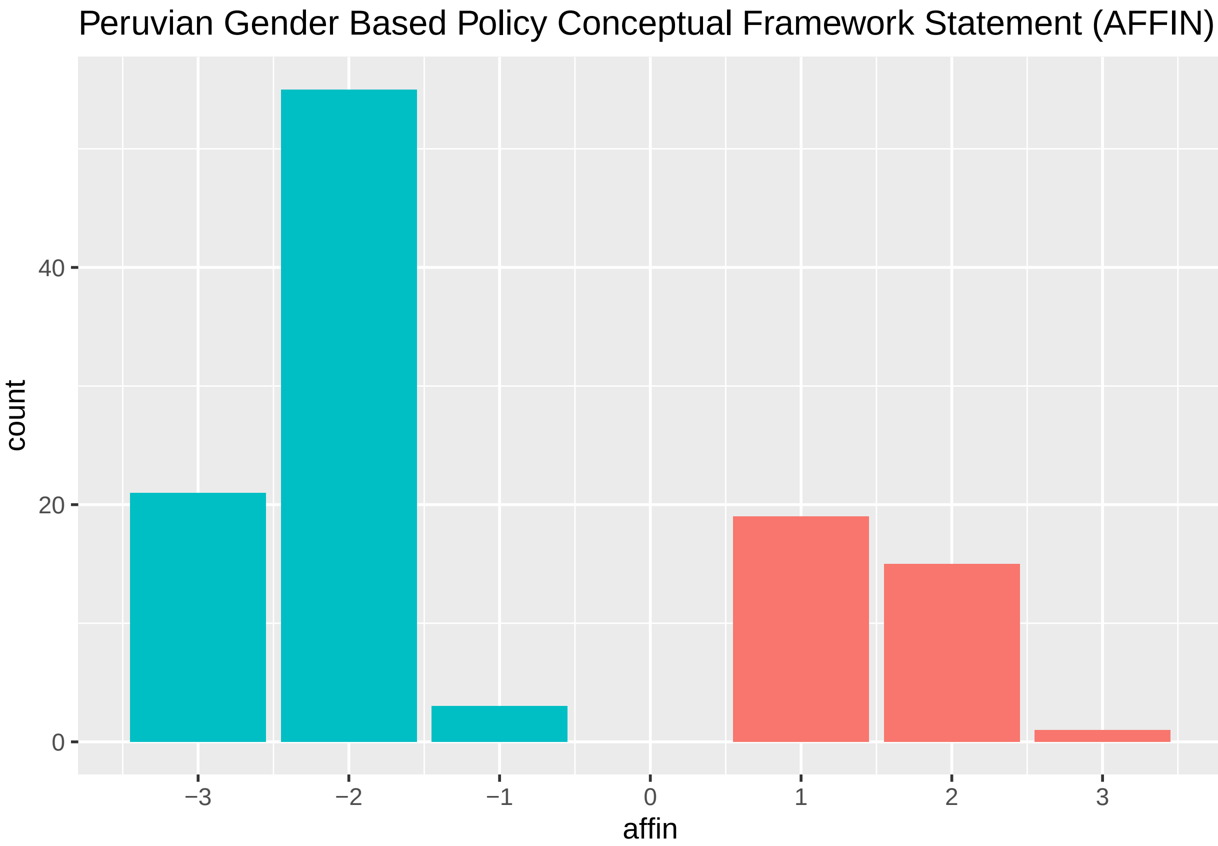


```
#ggsave("images/peru_sentiment_bing.png", width = 8, height = 7)
```

```
#Sentiment affin

ggplot(data = filter(peru_no_sw, !is.na(affin))) +
```

```r
  geom_histogram(aes(affin, fill = affin < 0), stat = "count") +
  scale_x_continuous(n.breaks = 7) +
  labs(title =
         "Peruvian Gender Based Policy Conceptual Framework Statement (AFFIN)") +
  theme(legend.position = "none")
```

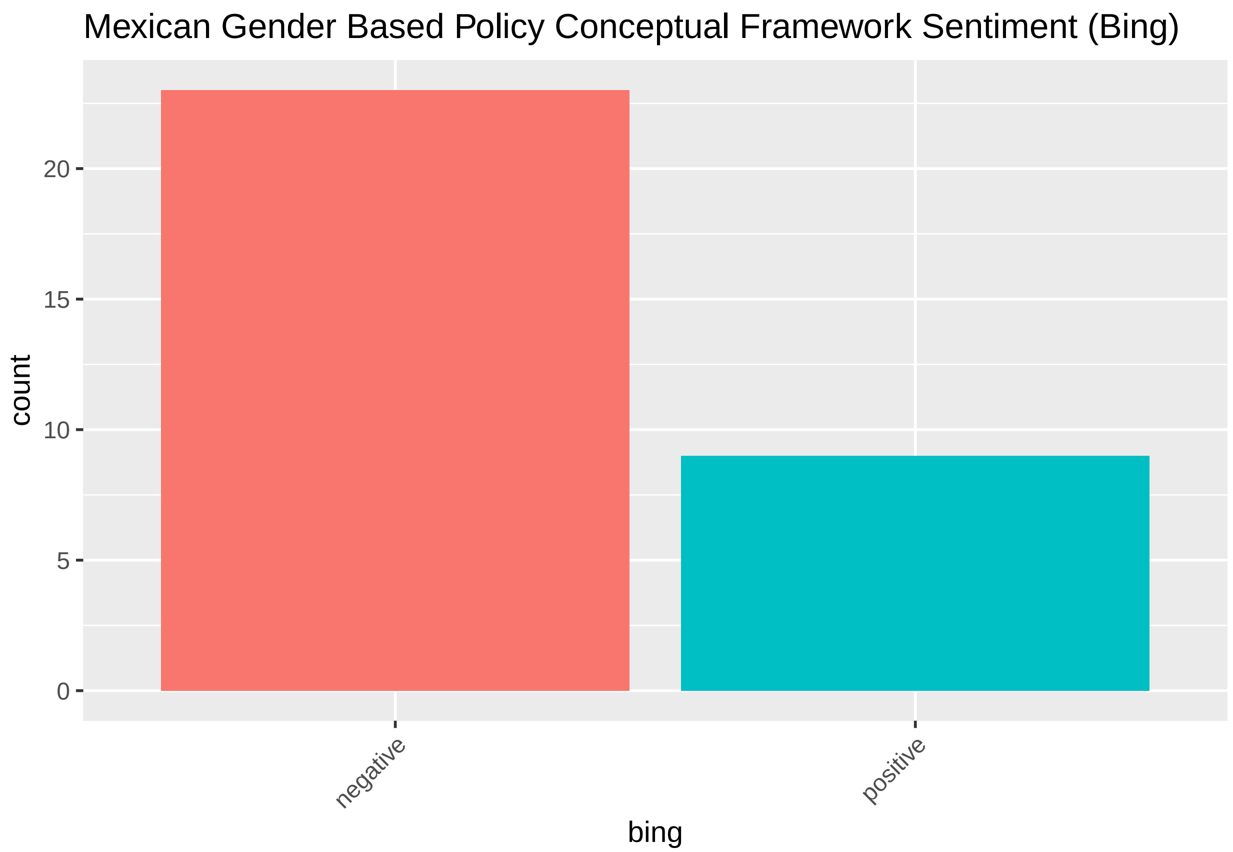
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```r
#ggsave("images/peru_sentiment_affin.png", width = 8, height = 7)
```

3.2 Plotting the Mexican policy document

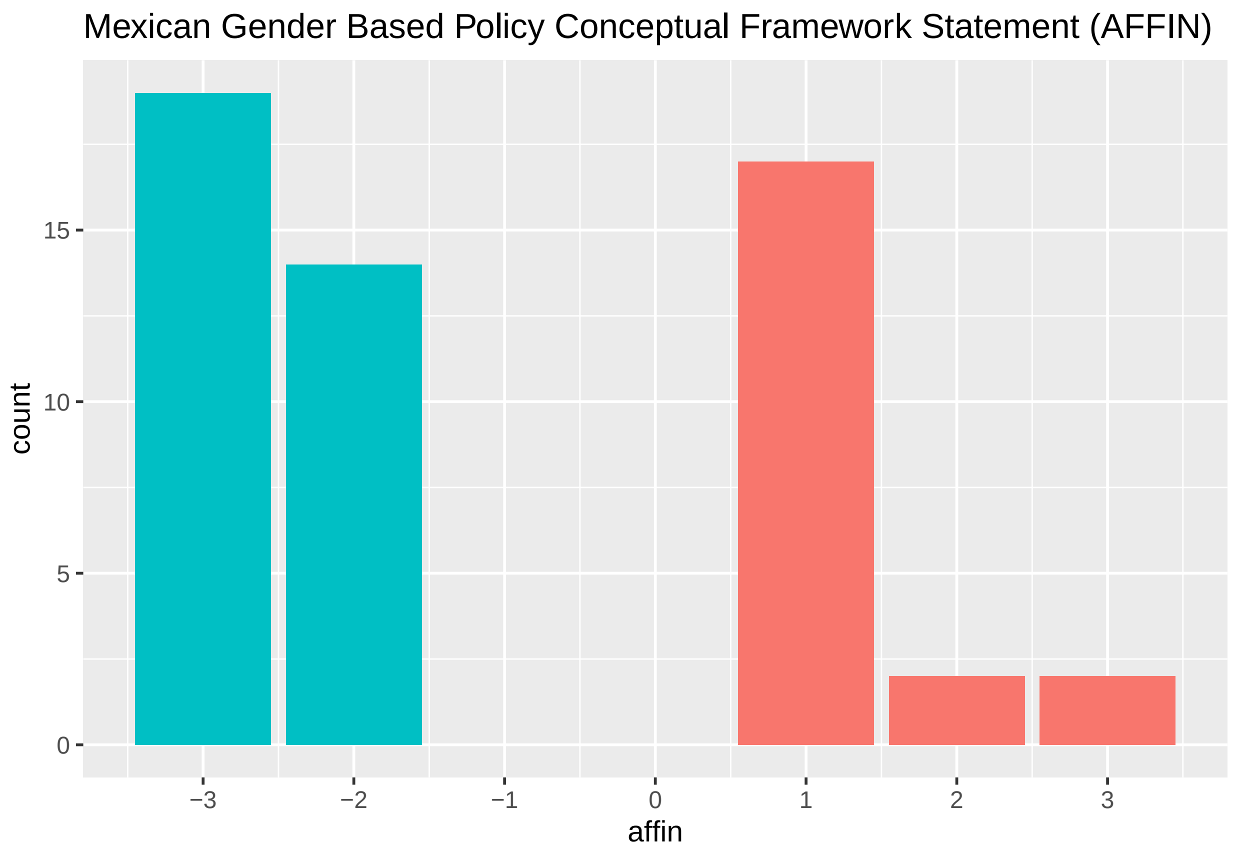
```r
# Plotting the Mexican policy document
mexico_no_sw <- mexico_no_sw %>%
  left_join(sentiment_nrc, by = c("word_tokens" = "word")) %>%
  left_join(sentiment_afinn, by = c("word_tokens" = "word")) %>%
  left_join(sentiment_bing, by = c("word_tokens" = "word"))

ggplot(data = filter(mexico_no_sw, !is.na(bing))) +
  geom_histogram(aes(bing, fill = bing), stat = "count") +
  scale_x_discrete(guide = guide_axis(angle = 45)) +
  labs(title = "Mexican Gender Based Policy Conceptual Framework Sentiment (Bing)") +
  theme(legend.position = "none")
```



```
#ggsave("images/mexico_sentiment_bing.png", width = 8, height = 7)
```

```
#Sentiment affin
ggplot(data = filter(mexico_no_sw, !is.na(affin))) +
  geom_histogram(aes(affin, fill = affin < 0), stat = "count") +
  scale_x_continuous(n.breaks = 7) +
  labs(title = "Mexican Gender Based Policy Conceptual Framework Statement (AFFIN)") +
  theme(legend.position = "none")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```
#ggsave("images/mexico_sentiment_affin.png", width = 8, height = 7)
```

3.3 Create wordclouds

4 Analysis Regressions

```
fit1 <- feols(formula(peru_violencia_fisica ~ Index | years_old + education_level
  + employed + marital_status + Departamento), peru_data_long)
```

```
## NOTE: 263 observations removed because of NA values (LHS: 263).
```

```
fit2 <- feols(formula(peru_violencia_sexual ~ Index | years_old + education_level
  + employed + marital_status + Departamento), peru_data_long)
```

```
## NOTE: 263 observations removed because of NA values (LHS: 263).
```

```
fit3 <- feols(formula(peru_violencia_psicologica ~ Index | years_old + education_level
  + employed + marital_status + Departamento), peru_data_long)
```

```
## NOTE: 263 observations removed because of NA values (LHS: 263).
```

```
etable <- etable(list(fit1, fit2,fit3),
  tex = FALSE,
  fitstat = c("n", "r2"), signif.code = NA
  #, file = paste0(path, "/images/reg_peru.txt")
)


etable
```

```
                              model 1               model 2
```

Dependent Var.: peru__violencia__fisica peru__violencia__sexual

Index 0.6302 (0.2447) 0.3041 (0.1173) Fixed-Effects: years__old Yes Yes education__level Yes Yes employed Yes Yes marital__status Yes Yes Departamento Yes Yes S.E.: Clustered by: years__old by: years__old Observations 2,575 2,575 R2 0.80550 0.74227

```
                                   model 3
```

Dependent Var.: peru__violencia__psicologica

Index 1.065 (0.4790) Fixed-Effects: years__old Yes education__level Yes employed Yes marital__status Yes Departamento Yes S.E.: Clustered by: years__old Observations 2,575 R2 0.45711

```
fit1 <- feols(formula(mexico_violencia_fisica ~ Index | education_level
    + employed + marital_status + state), mexico_data)
```

```
## NOTE: 22,909 observations removed because of NA values (LHS: 3,386, Fixed-effects: 19,523).
```

```
    fit2 <- feols(formula(mexico_violencia_sexual ~ Index | education_level
    + employed + marital_status + state), mexico_data)
```

```
## NOTE: 22,909 observations removed because of NA values (LHS: 3,386, Fixed-effects: 19,523).
```

```
    fit3 <- feols(formula(mexico_violencia_psicologica ~ Index | education_level
    + employed + marital_status + state), mexico_data)
```

```
## NOTE: 22,909 observations removed because of NA values (LHS: 3,386, Fixed-effects: 19,523).
```

```
etable <- etable(list(fit1, fit2,fit3),
  tex = FALSE,
  fitstat = c("n", "r2"), signif.code = NA
  #, file = paste0(path, "/images/reg_mex.txt")
)


etable
```

```
                            model 1                 model 2
```

Dependent Var.: mexico_violencia_fisica mexico_violencia_sexual

Index 1.101 (0.4256) 5.401 (2.006) Fixed-Effects: —————————— —————————— education_level Yes Yes employed Yes Yes marital_status Yes Yes state Yes Yes ______________________ ___________________________________ ___________________________________ S.E.: Clustered by: education_level by: education_level Observations 173,864 173,864 R2 0.80234 0.65061

```
                                 model 3
```

Dependent Var.: mexico_violencia_psicologica

Index 2.367 (0.8686) Fixed-Effects: ————————————- education_level Yes employed Yes marital_status Yes state Yes _______________________ ____ S.E.: Clustered by: education_level Observations 173,864 R2 0.77082