Limpieza de base de datos censo de población y vivienda (INEGI, 2020)

Katia Michelle Villarnobo

29/11/2021

Descripcción

Esta código se llevar acabo la limpieza de datos poblacional obtenida del censo de población y vivienda (INEGI, 2020)

Insumos: Se utilizó la información del Censo de Poblacion de vivienda 2020 en formato csv INEGI,2020. A la cual se le realizo un pre procesamiento, se realizó un filtro dee los municipios que estan en nuestra zona de estudio y se exporto como *"Poblacion_ISC.csv"*.

Limpieza de la base de datos

Para limpiar la base de datos se tuvieron que tomar en cuenta las siguientes consideraciones con relación a la información de censo de Población y vivienda 2020:

- 1. Los valores nulos estan expresados con *, hay que sustituir estos valores por ceros para poder realizar el análisis estadístico
  2. Las columnas que contienen información de población es de tipo caracter por lo que no se puedes procesar númericamente, por lo tanto hay que cambiar el tipo de dato a númerico.
  3. Los AGEB de cada municipio no tienen un valor unico por lo que hay que procesar cada municipio por separado para que al momento de unir lo con la geometría de la capa de AGEBS no se le asigne un valor erroneo a cada polígono
  4. La base de datos no solo contiene información por cada municipio si no que también tiene valores de población total, hay que filtrar los datos para que estos valores no afecten el análisis, el filtro se realizará NOM_LOC== "Total AGEB urbana" ya que es una característica única de estos reguistros de población total
  5. La información de la base de datos da la información de la población en función de cada manzana y no respecto a cada AGEB por lo que es importante agrupar los valores por AGEB y hacer la suma de la población total formando una base de datos para cada AGEB
  6. Para este analísis no son importantes los AGEB con poblacion total igual a cero por lo que hay que hacer un filtro de esos valores con el campo manzana =0
-

```r
### Bibliotecas

library(dplyr)
 library(ggplot2)
 library(readxl)

poblacion <- read.csv("GUANAJUATO.csv", header = TRUE )  ##IMPORTAMOS EL CSV

poblacion[poblacion=="*"] <- ""    ## Asignamos valores vacíos a los que tienen asteriscos

###cambiamos el tipo de dato de chr a num

 poblacion<- poblacion %>% as_tibble() %>% mutate(across(c(9:ncol(.)), as.numeric))

 poblacion<-transform(poblacion, MUN= as.character(MUN))
 poblacion<-transform(poblacion, LOC= as.character(LOC))

poblacion[is.na(poblacion)] <- 0 # A los valores con NA les asignamos el valor de cero
#str(poblacion)

#names(poblacion)

###############    Campo para el join     ################


cols <- c("ENTIDAD", "MUN", "LOC", "AGEB")
#colocamos el nombre de las columnas a unir

poblacion<- rename(poblacion, ENTIDAD = ï..ENTIDAD)
#Hacemos un campo llamado clave que genere la concatenación de campos  separados por _
poblacion$Clave <- apply(poblacion[ ,cols],1 , paste, collapse ="_"  )

#names(poblacion)  # revisamos que el campo fue creado



#############   Array de Poblacion:total ##################


#Hacemos un filtro solo para los valores de población total

poblacion_T<- poblacion %>% filter(NOM_LOC== "Total AGEB urbana")



##########    Array número de manzanas  ###############

# Vamos a hacer un filtro para aquellos datos cuya manzana sea diferente de cero
# Esto nos eliminara los registros de poblacion total y nos dejará solo los valores por
#manzana

N_MANZANAS<- poblacion %>% filter(MZA !=0)
```

```r
#Vamos a considerar solo los valores de aquellas manzanas que tengan una población
#diferente de cero
#N_MANZANAS<- N_MANZANAS%>% filter(POBTOT!=0)

#View(N_MANZANAS
## Hacemos un análisis por AGEB usando group by Calculamos solo el número de manzanas
#involucradas
#names(N_MANZANAS)


n<-N_MANZANAS %>%  group_by(Clave) %>% summarise( N_MZ = n())



#str(pobtotal)

##############     Obtener array final     ##############


POB_F<- merge(n,poblacion_T)



str(POB_F)
```

```
## 'data.frame':    3674 obs. of  232 variables:
##  $ Clave      : chr  "11_1_1_0059" "11_1_1_0063" "11_1_1_0129" "11_1_1_0133" ...
##  $ N_MZ       : int  38 45 22 54 40 10 85 31 1 9 ...
##  $ ENTIDAD    : int  11 11 11 11 11 11 11 11 11 11 ...
##  $ NOM_ENT    : chr  "Guanajuato" "Guanajuato" "Guanajuato" "Guanajuato" ...
##  $ MUN        : chr  "1" "1" "1" "1" ...
##  $ NOM_MUN    : chr  "Abasolo" "Abasolo" "Abasolo" "Abasolo" ...
##  $ LOC        : chr  "1" "1" "1" "1" ...
##  $ NOM_LOC    : chr  "Total AGEB urbana" "Total AGEB urbana" "Total AGEB urbana" "Total AGEB urbana"
##  $ AGEB       : chr  "0059" "0063" "0129" "0133" ...
##  $ MZA        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ POBTOT     : num  3085 3749 1617 3723 2379 ...
##  $ POBFEM     : num  1643 1987 806 1903 1246 ...
##  $ POBMAS     : num  1442 1762 811 1820 1133 ...
##  $ P_0A2      : num  130 142 84 169 102 48 367 104 0 51 ...
##  $ P_0A2_F    : num  73 71 38 77 55 24 175 52 0 23 ...
##  $ P_0A2_M    : num  57 71 46 92 47 24 192 52 0 28 ...
##  $ P_3YMAS    : num  2955 3607 1533 3554 2273 ...
##  $ P_3YMAS_F  : num  1570 1916 768 1826 1189 ...
##  $ P_3YMAS_M  : num  1385 1691 765 1728 1084 ...
##  $ P_5YMAS    : num  2882 3536 1476 3448 2199 ...
##  $ P_5YMAS_F  : num  1537 1876 740 1777 1151 ...
##  $ P_5YMAS_M  : num  1345 1660 736 1671 1048 ...
##  $ P_12YMAS   : num  2577 3160 1251 2974 1954 ...
##  $ P_12YMAS_F : num  1389 1695 634 1526 1030 ...
##  $ P_12YMAS_M : num  1188 1465 617 1448 924 ...
##  $ P_15YMAS   : num  2442 2983 1180 2769 1842 ...
##  $ P_15YMAS_F : num  1316 1604 602 1420 976 ...
```

```
##  $ P_15YMAS_M : num  1126 1379 578 1349 866 ...
##  $ P_18YMAS   : num  2311 2815 1091 2546 1726 ...
##  $ P_18YMAS_F : num  1248 1516 559 1316 920 ...
##  $ P_18YMAS_M : num  1063 1299 532 1230 806 ...
##  $ P_3A5      : num  122 120 80 173 106 40 337 123 6 48 ...
##  $ P_3A5_F    : num  62 60 39 84 55 23 164 66 4 17 ...
##  $ P_3A5_M    : num  60 60 41 89 51 17 173 57 0 31 ...
##  $ P_6A11     : num  256 327 202 407 213 95 714 206 5 110 ...
##  $ P_6A11_F   : num  119 161 95 216 104 54 361 101 0 53 ...
##  $ P_6A11_M   : num  137 166 107 191 109 41 353 105 3 57 ...
##  $ P_8A14     : num  305 407 205 471 263 110 886 252 9 137 ...
##  $ P_8A14_F   : num  152 207 100 247 128 64 444 114 4 62 ...
##  $ P_8A14_M   : num  153 200 105 224 135 46 442 138 5 75 ...
##  $ P_12A14    : num  135 177 71 205 112 45 394 112 4 54 ...
##  $ P_12A14_F  : num  73 91 32 106 54 27 187 46 0 24 ...
##  $ P_12A14_M  : num  62 86 39 99 58 18 207 66 0 30 ...
##  $ P_15A17    : num  131 168 89 223 116 43 405 78 4 55 ...
##  $ P_15A17_F  : num  68 88 43 104 56 23 200 35 0 28 ...
##  $ P_15A17_M  : num  63 80 46 119 60 20 205 43 0 27 ...
##  $ P_18A24    : num  341 446 180 451 298 98 861 257 9 119 ...
##  $ P_18A24_F  : num  168 223 89 203 135 52 421 123 8 56 ...
##  $ P_18A24_M  : num  173 223 91 248 163 46 440 134 0 63 ...
##  $ P_15A49_F  : num  791 982 417 997 623 ...
##  $ P_60YMAS   : num  587 710 176 399 361 79 366 191 29 122 ...
##  $ P_60YMAS_F : num  335 403 84 209 202 43 167 103 16 62 ...
##  $ P_60YMAS_M : num  252 307 92 190 159 36 199 88 13 60 ...
##  $ REL_H_M    : num  87.8 88.7 100.6 95.6 90.9 ...
##  $ POB0_14    : num  643 766 437 954 533 ...
##  $ POB15_64   : num  1995 2472 1075 2527 1597 ...
##  $ POB65_MAS  : num  447 511 105 242 245 57 209 128 24 80 ...
##  $ PROM_HNV   : num  2.2 2.19 2.5 2.16 2.27 2.36 2.17 2.59 3.22 2.36 ...
##  $ PNACENT    : num  2824 3550 1520 3490 2225 ...
##  $ PNACENT_F  : num  1505 1883 754 1786 1164 ...
##  $ PNACENT_M  : num  1319 1667 766 1704 1061 ...
##  $ PNACOE     : num  240 191 88 219 138 11 257 75 10 29 ...
##  $ PNACOE_F   : num  125 100 45 112 72 8 116 39 5 16 ...
##  $ PNACOE_M   : num  115 91 43 107 66 3 141 36 5 13 ...
##  $ PRES2015   : num  2782 3490 1440 3392 2180 ...
##  $ PRES2015_F : num  1493 1850 724 1749 1143 ...
##  $ PRES2015_M : num  1289 1640 716 1643 1037 ...
##  $ PRESOE15   : num  94 40 28 53 14 5 58 25 0 5 ...
##  $ PRESOE15_F : num  40 24 14 27 7 0 23 14 0 3 ...
##  $ PRESOE15_M : num  54 16 14 26 7 3 35 11 0 0 ...
##  $ P3YM_HLI   : num  12 0 10 12 0 0 3 15 4 0 ...
##  $ P3YM_HLI_F : num  5 0 4 4 0 0 0 6 0 0 ...
##  $ P3YM_HLI_M : num  7 0 6 8 0 0 0 9 3 0 ...
##  $ P3HLINHE   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ P3HLINHE_F : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ P3HLINHE_M : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ P3HLI_HE   : num  12 0 9 12 0 0 3 15 4 0 ...
##  $ P3HLI_HE_F : num  5 0 3 4 0 0 0 6 0 0 ...
##  $ P3HLI_HE_M : num  7 0 6 8 0 0 0 9 3 0 ...
##  $ P5_HLI     : num  12 0 8 12 0 0 3 15 4 0 ...
##  $ P5_HLI_NHE : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ P5_HLI_HE  : num  12 0 8 12 0 0 3 15 4 0 ...
##  $ PHOG_IND   : num  21 0 10 29 0 0 7 31 4 0 ...
##  $ POB_AFRO   : num  38 14 11 12 47 3 13 13 0 0 ...
##  $ POB_AFRO_F : num  19 8 3 5 24 0 6 5 0 0 ...
##  $ POB_AFRO_M : num  19 6 8 7 23 0 7 8 0 0 ...
##  $ PCON_DISC  : num  245 231 86 169 149 31 313 137 9 53 ...
##  $ PCDISC_MOT : num  162 154 58 106 79 14 176 94 7 34 ...
##  $ PCDISC_VIS : num  67 75 18 43 67 11 120 50 0 10 ...
##  $ PCDISC_LENG: num  23 29 21 31 24 0 32 7 0 12 ...
##  $ PCDISC_AUD : num  20 41 8 24 27 3 41 16 0 6 ...
##  $ PCDISC_MOT2: num  48 43 12 43 23 3 27 18 0 8 ...
##  $ PCDISC_MEN : num  37 42 14 32 29 3 40 18 0 6 ...
##  $ PCON_LIMI  : num  497 580 300 349 425 74 465 289 9 93 ...
##  $ PCLIM_CSB  : num  234 290 119 155 138 20 187 110 6 52 ...
##  $ PCLIM_VIS  : num  266 346 190 171 295 55 259 212 6 18 ...
##  $ PCLIM_HACO : num  19 37 13 28 47 3 24 11 5 8 ...
##  $ PCLIM_OAUD : num  71 137 33 67 95 12 57 36 6 21 ...
##  $ PCLIM_MOT2 : num  35 44 20 22 35 0 9 4 5 0 ...
##   [list output truncated]
```

```
##Exportar csv

write.csv(POB_F, file="POBLACION_AGEB.csv")
```