Metodología

Bryan Castillo

17 de junio de 2021

Metodología

El método consite basicamente en obtener las bases de datos en formato comma-separated *values* (csv) a escala de personas de los censos 1992, 2002 y 2017 desde la página de internet del *Instituto Nacional de Estadísticas* (INE). Debido al gran esfuerzo de procesamiento que requiere procesar la gran cantidad de dato adquiridos, los archivos de datos censales para los tres periodos fueron cargados en el gestor de base de datos *PostgreSQL* a través de una librería, *RpostgreSQL*, que sirve como intérprete entre los lenguajes de *R* y *SQL* y también como interfaz entre el **IDE** *RStudio* y el gestor de bases de datos *PostgreSQL*, de esta manera, además de cargar las bases de datos se pudieron elaborar las consultas pertinentes a los datos con ayuda de la librería *dbplyr*.

Con ayuda de las librerías mencionadas anteriormente se realizó la siguiente secuencia de pasos:

En primer lugar se configura y establece la conexión entre el lenguaje R y su interfaz gráfica RStudio con el gestor de base de datos PostgreSQL

```
fun_connect<-function(){dbConnect(RPostgres::Postgres(),
                                  dbname = 'censos',
                                  host = 'localhost',
                                  port = 5432, # or any other port specified by your DBA
                                  user = 'usuario',
                                  password = 'contrasena',
                                  options="-c search_path=censos")}

conn <- fun_connect()
```

Luego se procede a crear tablas a partir de las características de los datos en formato *csv* es importante conocer el tipo de datos que se cargará, el número de filas, columnas y el tamaño de la celda que ocupara cada dígito o caracter, para esto se creo una función que hace un resumen del número máximo de carácteres de cada fila y columna de los datos a usar.

Devuelve el tamaño máximo de caraácteres o dígitos que debe tener la variable:

```
#Ref:grasshoppermouse: https://www.reddit.com/r/rstats/comments/azumgy/is_there_a_summarylike_function_

test2002<-summarise_all(readRDS("C:/CEDEUS/Censos/Censo2002_Persona_Full.Rds"),funs(
  case_when(is.character(.)~max(nchar(.), na.rm=T),
            is.numeric(.) ~max(nchar(as.character(.)), na.rm=T))))

test1992<-summarise_all(readRDS("C:/CEDEUS/Censos/Censo1992_Persona_Full.Rds"),funs(
  case_when(is.character(.)~max(nchar(.), na.rm=T),
            is.numeric(.) ~max(nchar(as.character(.)), na.rm=T))))

test2017<-summarise_all(readRDS("C:/CEDEUS/Censos/Censo2017_Persona_Full.Rds"),funs(
```

```
  case_when(is.character(.)~max(nchar(.), na.rm=T),
            is.numeric(.) ~max(nchar(as.character(.)), na.rm=T))))
```

Devuelve la clase de la variable que se debe usar para crear la tabla:

```
abreColumnas<-function(x){tidyr::spread(dplyr::group_by(as.data.frame(summary.default(readRDS(x),funs(su

censo1992<-abreColumnas("C:/CEDEUS/Censos/Censo1992_Persona_Full.Rds")
censo2002<-abreColumnas("C:/CEDEUS/Censos/Censo2002_Persona_Full.Rds")
censo2017<-abreColumnas("C:/CEDEUS/Censos/Censo2017_Persona_Full.Rds")
```

Creo una función que mezcla los dos output anteriores y da los parámetros para crear la tabla:

```
tablita<-function(x,y){
  a<-cbind(variables=row.names(as.data.frame(t(x))),as.data.frame(t(x)))

  rownames(a)<- 1:nrow(a)

  a<-left_join(y, a, by=c("Var1"="variables"))
  a<-a[,c(1,4,5)]
  colnames(a)<-c("variable","clase","largo")
  return(a)
}

censo_1992<-tablita(test1992, censo1992)
censo_2002<-tablita(test2002, censo2002)
censo_2017<-tablita(test2017, censo2017)
```

Creo esquema y me conecta a la base de datos:

```
#dbSendQuery(conn, "CREATE SCHEMA censos")

dbSendQuery(conn, "set search_path to censos;")
```

Creo una función que me construye una consulta para crear las tablas con los datos generados a partir de la función tablita:

```
creaTabla<-function(x){
  y<-deparse(substitute(x))

  x$variable<-gsub("\\.","",x$variable)
  x$variable<-stringr::str_replace_all(x$variable,"ñ","n")
  x<-paste(c(paste("ID", "SERIAL PRIMARY KEY"),
             paste(
               paste(x$variable, x$clase),
               "(",x$largo,")", "NOT NULL")),collapse=",")

                    x<-paste0("CREATE TABLE ",y," (",x, ")")
                    return(x)}

dbSendQuery(conn,creaTabla(censo_1992))
dbSendQuery(conn,creaTabla(censo_2002))
dbSendQuery(conn,creaTabla(censo_2017))
```

Luego de crear las tablas creo una consulta que las llena:

```
dbSendQuery(conn, paste("copy censo_1992",creaCopy(censo_1992)," from PROGRAM '7z e -so C:/CEDEUS/2021/a
dbSendQuery(conn, paste("copy censo_2002",creaCopy(censo_2002)," from PROGRAM '7z e -so C:/CEDEUS/2021/a
dbSendQuery(conn, paste("copy censo_2012",creaCopy(censo_2012)," from PROGRAM '7z e -so C:/CEDEUS/2021/a
dbSendQuery(conn, paste("copy censo_2017",creaCopy(censo_2017)," from PROGRAM '7z e -so C:/CEDEUS/2021/a
```

Una vez cargados los datos en el gestor de base de datos *PostgreSQL* empezamos a realizar consultas, en específico se quiere saber si las personas adultos mayores que habitan en la zona de estudio al momento de ser censadas residían en esa comuna o no, y si no preguntar de dónde venían.

Se procedió a crear tablas dónde se pone el número y porcentaje de personas >=60 años que vivían en otra comuna al momento de ser encuestadas y a crear *"heatmaps"* y *"alluvial diagrams"* para sintetizar la información que nos dice de que comuna vienen las personas que hace 5 años vivían en otra comuna.