practica_clase2

guada

24/2/2020

Practica 2: Aprendiendo Tidyverse con dataset de victimas del terrorismo de estado

Actividades:

1. Subir a R el dataset de "Listado de víctimas del accionar represivo ilegal" elaborado por el Ministerio de Justicia y Derechos Humanos (https://datos.gob.ar/dataset/justicia-registro-unificado-victimas-terrorismo-estado--ruvte-) que contiene un listado de las desapariciones y asesinatos ocurridos durante 1966 y 1983

(Recomendaciones: Como podemos ver el dataset tiene inconvenientes con los caracteres especiales, eso significa que tenemos que cambiarle el encoding.. el mismo se hace agregando dentro de la funcion *read.csv(path = , encoding = "Latin-1")*)

```
data_terrorismo_arg <- read.csv("http://datos.jus.gob.ar/dataset/d43fa140-f43f-4cc2-8491-b1d8bb899de4/r
```

2. Realizamos un head() para ver las variables

```
head(data_terrorismo_arg)
```

```
##   anio_denuncia   tipificacion_ruvte id_unico_ruvte    apellido_paterno_nombres
## 1          1984 DESAPARICION FORZADA        ID 5389       ABACHIAN  JUAN CARLOS
## 2          1984 DESAPARICION FORZADA          ID 87          ABAD  ANA CATALINA
## 3          1984 DESAPARICION FORZADA       ID 11788         ABAD  JULIO RICARDO
## 4          1984            ASESINATO        ID 9907         ABAD  OSCAR GERARDO
## 5          1984 DESAPARICION FORZADA          ID 89 ABAD  ROBERTO RODOLFO TOMAS
## 6          1984 DESAPARICION FORZADA          ID 88      ABAD  ROUMALDO RICARDO
##   apellido_materno apellido_casada edad_al_momento_del_hecho   documentos
## 1       BEDROSSIAN                                  26 años   LE 8293245
## 2         SCARLATA         PERUCCA                  24 años  LC 10048122
## 3          CORONEL                                  21 años DNI 10283544
## 4           DOMATO                                  25 años DNI 10353245
## 5           ZABALA                                  23 años DNI 10650064
## 6            ARAYA                                  54 años   LE 3498462
##   anio_nacimiento provincia_nacimiento pais_nacimiento nacionalidad embarazo
## 1            1950         BUENOS AIRES       ARGENTINA    ARGENTINA
## 2            1951              MENDOZA       ARGENTINA    ARGENTINA
## 3            1954              TUCUMAN       ARGENTINA    ARGENTINA
```

```
## 4            1951         BUENOS AIRES       ARGENTINA    ARGENTINA
## 5            1953      CAPITAL FEDERAL       ARGENTINA    ARGENTINA
## 6            1921              TUCUMAN       ARGENTINA    ARGENTINA
##                    fecha_lugar_detencion_secuestro
## 1               26/12/1976   LA PLATA  BUENOS AIRES
## 2            15/08/1976   CORDOBA  CAPITAL  CORDOBA
## 3                        NOV/1976   CAPITAL FEDERAL
## 4               08/10/1976   LA PLATA  BUENOS AIRES
## 5 09/08/1976   FLORIDA  VICENTE LOPEZ  BUENOS AIRES
## 6       04/05/1975   SANTA LUCIA  MONTEROS  TUCUMAN
##                              fecha_lugar_asesinato_o_hallazgo_de_restos
## 1                                                                   ---
## 2                                                                   ---
## 3                                                                   ---
## 4 21/10/1976   GRAL. MANSILLA (BARTOLOME BAVIO)  MAGDALENA  BUENOS AIRES
## 5                                                                   ---
## 6                                                                   ---
##   fotografia provincia_nacimiento_indec_id pais_nacimiento_indec_id
## 1         Sí                             6                      ARG
## 2         Sí                            50                      ARG
## 3         No                            90                      ARG
## 4         No                             6                      ARG
## 5         Sí                             2                      ARG
## 6         No                            90                      ARG
```

3. ¿Qué tipo de variables tiene la tabla? (numericas, caracteres, etc)

```
typeof(data_terrorismo_arg$anio_nacimiento)
```

```
## [1] "character"
```

4. Carguemos la librería de tidyverse y conozcamos más sobre la información que brinda el dataset

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1
```

```
## <U+2713> ggplot2 3.2.1     <U+2713> purrr   0.3.3
## <U+2713> tibble  2.1.3     <U+2713> dplyr   0.8.3
## <U+2713> tidyr   1.0.0     <U+2713> stringr 1.4.0
## <U+2713> readr   1.3.1     <U+2713> forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
## Warning: package 'tibble' was built under R version 3.6.2
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## Warning: package 'stringr' was built under R version 3.6.2

## Warning: package 'forcats' was built under R version 3.6.2

## -- Conflicts ------------------------------------------------------------------------ tidyverse_conflic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

a. Queremos saber, en principio, cuál es el país con mayor cantidad de desapariciones forzadas según la nacionalidad.

```
data_terrorismo_arg %>%
  filter(tipificacion_ruvte == "DESAPARICION FORZADA" |
           tipificacion_ruvte == "DESAPARICION FORZADA / PROBADO EL DECESO" |
           tipificacion_ruvte == "DESAPARICION FORZADA / EXHUMADOS E IDENTIFICADOS SUS RESTOS" |
           tipificacion_ruvte == "DESAPARICION FORZADA / INVESTIGADO EN CAUSA JUDICIAL" |
           tipificacion_ruvte == "DESAPARICION FORZADA / EN INVESTIGACION" |
          tipificacion_ruvte ==  "DESAPARICION FORZADA / A DETERMINAR TIPIFICACION" |
           tipificacion_ruvte ==  "DESAPARICION FORZADA (NIÑA) / EXHUMADOS E IDENTIFICADOS SUS RESTOS"
           tipificacion_ruvte ==  "DESAPARICION FORZADA (NIÑO) / EXHUMADOS E IDENTIFICADOS SUS RESTOS"
  group_by(pais_nacimiento) %>%
  summarize(cantidad = n()) %>%
  arrange(desc(cantidad))
```

```
## # A tibble: 37 x 2
##    pais_nacimiento cantidad
##    <chr>              <int>
##  1 ARGENTINA           6053
##  2 sin datos            236
##  3 URUGUAY              125
##  4 PARAGUAY              73
##  5 CHILE                 71
##  6 ITALIA                48
##  7 ESPAÑA                46
##  8 BOLIVIA               28
##  9 PERU                  15
## 10 BRASIL                 9
## # ... with 27 more rows
```

a.a. ¿Cuál es la provincia con mayor cantidad de desapariciones forzadas?

```
data_terrorismo_arg %>%
  filter(tipificacion_ruvte == "DESAPARICION FORZADA" |
           tipificacion_ruvte == "DESAPARICION FORZADA / PROBADO EL DECESO" |
           tipificacion_ruvte == "DESAPARICION FORZADA / EXHUMADOS E IDENTIFICADOS SUS RESTOS" |
           tipificacion_ruvte == "DESAPARICION FORZADA / INVESTIGADO EN CAUSA JUDICIAL" |
           tipificacion_ruvte == "DESAPARICION FORZADA / EN INVESTIGACION" |
```

```
          tipificacion_ruvte ==  "DESAPARICION FORZADA / A DETERMINAR TIPIFICACION" |
           tipificacion_ruvte ==  "DESAPARICION FORZADA (NIÑA) / EXHUMADOS E IDENTIFICADOS SUS RESTOS" |
           tipificacion_ruvte ==  "DESAPARICION FORZADA (NIÑO) / EXHUMADOS E IDENTIFICADOS SUS RESTOS"
  group_by(provincia_nacimiento) %>%
  summarize(cantidad = n()) %>%
  arrange(desc(cantidad))
```

```
## # A tibble: 26 x 2
##    provincia_nacimiento  cantidad
##    <chr>                    <int>
##  1 "CAPITAL FEDERAL"         1737
##  2 "BUENOS AIRES"            1578
##  3 "TUCUMAN"                  598
##  4 "CORDOBA"                  528
##  5 ""                         464
##  6 "SANTA FE"                 388
##  7 "sin datos"                236
##  8 "ENTRE RIOS"               209
##  9 "MENDOZA"                  144
## 10 "SANTIAGO DEL ESTERO"      122
## # ... with 16 more rows
```

b. Queremos saber cuántos años tendrian al dia de hoy las personas registradas en el dataset cuya provincia de nacimiento sea "BUENOS AIRES" y que figure en el nuevo dataset sólo las columnas de "provincia_nacimiento", "anio_nacimiento" y la "edad_actual".

```
#Primero vamos a tener que pasarlo a numerico
data_terrorismo_arg$anio_nacimiento <- as.numeric(as.character(data_terrorismo_arg$anio_nacimiento))
```

```
## Warning: NAs introducidos por coerción
```

```
data_edad <- data_terrorismo_arg %>%
  filter(provincia_nacimiento == "BUENOS AIRES") %>%
  mutate(edad_actual = 2019-anio_nacimiento) %>%
  select(provincia_nacimiento, anio_nacimiento, edad_actual) %>%
  drop_na()
```

b.a.¿Cuál es la edad promedio que tendrian al día de hoy?

```
mean(data_edad$edad_actual)
```

```
## [1] 70.46383
```

5. ¿Qué pasa si queremos saber el género de la persona y no figura en el dataset? Vamos paso a paso.

a. Podemos empezar construirlo gracias al registro realizado por el gobierno de la Ciudad que tendremos que cargar en nuestro environment y que esta cargado en el github. (Acordemosnos de modificar el encoding, que en este caso es: "UTF-8").

```
nombres <- read.csv("https://raw.githubusercontent.com/Guadag12/R4RRII/master/Clase%202/nombres.csv", en
```

b. La idea es realizar un join y unir ambas tablas, sin embargo no tenemos una columna en nuestro dataset que unicamente sean nombres. Hay que construirla con la funcion "cSplit()" del paquete splitstackshape

```
library(splitstackshape)
```

```
## Warning: package 'splitstackshape' was built under R version 3.6.2
```

```
data_terrorismo_arg <- cSplit(indt = data_terrorismo_arg, splitCols = 'apellido_paterno_nombres',  sep =

#llamamos nombre a la columna que nos interesa
names(data_terrorismo_arg)[19] <- "nombre"

#vamos a trabajar con las columnas que nos interesan
data_terrorismo_arg <- data_terrorismo_arg %>% select(anio_denuncia, tipificacion_ruvte,id_unico_ruvte,a
documentos, anio_nacimiento,  provincia_nacimiento, pais_nacimiento, nacionalidad, embarazo,
fecha_lugar_detencion_secuestro, fecha_lugar_asesinato_o_hallazgo_de_restos, fotografia,
provincia_nacimiento_indec_id, pais_nacimiento_indec_id, apellido_paterno_nombres_1,
nombre )
```

c. Realizamos el join y eliminamos los duplicados!

```
data_terrorismo_arg1 <- left_join(data_terrorismo_arg, nombres, by = "nombre")
```

```
## Warning: Column `nombre` joining factors with different levels, coercing to
## character vector
```

```
#eliminamos las columnas que estan de mas
data_terrorismo_arg1 <- data_terrorismo_arg1[!duplicated(data_terrorismo_arg1$id_unico_ruvte), ]
```

¿Podemos conocer el porcentaje de personas desaparecidas segun el género?

```
data_terrorismo_arg1 %>%
  group_by(genero) %>%
  summarize(porcentaje = n()/8753)
```

```
## Warning: Factor `genero` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 3 x 2
##   genero porcentaje
##   <fct>       <dbl>
## 1 F           0.247
## 2 M           0.684
## 3 <NA>        0.0689
```

```r
#incluso si lo queremos hacer más prolijo seria algo asi:
data_terrorismo_arg1 %>%  group_by(genero) %>% summarize(porcentaje = paste0(round((n()/8753)*100, 2),
```

```
## Warning: Factor `genero` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
## # A tibble: 3 x 2
##   genero porcentaje
##   <fct>  <chr>
## 1 F      24.7%
## 2 M      68.41%
## 3 <NA>   6.89%
```