Sesion 3

Juvenal Campos

9/19/2019

Leer una base de datos.

En este caso, vamos a leer una base de datos que está almacenada desde internet. Para esto utilizaremos la función `read_delim` de la librería `readr`.

```
library(readr)

carpetas <-  read_delim("http://segasi.com.mx/clases/cide/datos/carpetas-de-investigacion-pgj-de-la-ciu
    ";",
    escape_double = FALSE,
    trim_ws = TRUE, skip = 6)
```

```
## Parsed with column specification:
## cols(
##   ao_hechos = col_double(),
##   mes_hechos = col_character(),
##   fecha_hechos = col_datetime(format = ""),
##   delito = col_character(),
##   categoria_delito = col_character(),
##   fiscalia = col_character(),
##   agencia = col_character(),
##   unidad_investigacion = col_character(),
##   alcaldia_hechos = col_character(),
##   colonia_hechos = col_character(),
##   ao_inicio = col_double(),
##   mes_inicio = col_character(),
##   fecha_inicio = col_datetime(format = ""),
##   calle_hechos = col_character(),
##   calle_hechos2 = col_character(),
##   longitud = col_double(),
##   latitud = col_double(),
##   geopoint = col_character()
## )
```

Explorando la base de datos

Exploramos la base de datos. Para esto, utilizamos las funciones `dim()` y `glimpse`.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.2
## v tidyr   0.8.3     v stringr 1.4.0
## v ggplot2 3.2.0     v forcats 0.4.0
```

```
## -- Conflicts ---------------------- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
class(carpetas)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
# Funcion para conocer las dimensiones de la base de datos
dim(carpetas)
```

```
## [1] 16623    18
```

```
# Para dar un vistazo a la base de datos
glimpse(carpetas)
```

```
## Observations: 16,623
## Variables: 18
## $ ao_hechos            <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019,...
## $ mes_hechos           <chr> "Agosto", "Agosto", "Agosto", "Agosto", "...
## $ fecha_hechos         <dttm> 2019-08-03 17:40:00, 2019-08-03 15:42:00...
## $ delito               <chr> "ABUSO SEXUAL", "ROBO A REPARTIDOR CON VI...
## $ categoria_delito     <chr> "DELITO DE BAJO IMPACTO", "ROBO A REPARTI...
## $ fiscalia             <chr> "INVESTIGACIÓN PARA LA ATENCIÓN DE DELITO...
## $ agencia              <chr> "FDS-2", "GAM-6", "XO-2", "TLP-4", "AO-4"...
## $ unidad_investigacion <chr> "FDS-2-02", "UI-2CD", "UI-1SD", "UI-2CD",...
## $ alcaldia_hechos      <chr> "TLALPAN", "GUSTAVO A MADERO", "XOCHIMILC...
## $ colonia_hechos       <chr> "CUMBRES DE TEPETONGO", "VALLEJO", "SANTA...
## $ ao_inicio            <dbl> 2019, 2019, 2019, 2019, 2019, 2019, 2019,...
## $ mes_inicio           <chr> "Agosto", "Agosto", "Agosto", "Agosto", "...
## $ fecha_inicio         <dttm> 2019-08-03 19:46:00, 2019-08-03 19:53:00...
## $ calle_hechos         <chr> "NO PRECISA CALLES", "SHUMMAN", "CARRETER...
## $ calle_hechos2        <chr> NA, "SAINT SAENZ", "DESIDERIO PEÑA", "ANI...
## $ longitud             <dbl> -99.18821, -99.13494, -99.08921, -99.1389...
## $ latitud              <dbl> 19.27284, 19.46920, 19.24722, 19.28201, 1...
## $ geopoint             <chr> "19.2728434436,-99.1882139267", "19.46920...
```

Limpieza de datos.

¿En qué consiste limpiar una Base de datos?

Consiste en pasar de una base cruda a una base que tenga exactamente la información que queremos.

```
# Limpiar base de datos.
# Queremos las 5 calles mas peligrosas de la ciudad de Mexico.

######################
# Son equivalentes. #
#####################

# Con el argumento sort desde la funcion count
carpetas %>%
  count(calle_hechos, sort = TRUE)
```

```
## # A tibble: 9,650 x 2
##    calle_hechos                  n
##    <chr>                     <int>
##  1 CALZADA DE TLALPAN           65
```

```
##  2 CALZADA DE GUADALUPE           63
##  3 AVENIDA TLAHUAC                62
##  4 CALZADA IGNACIO ZARAGOZA       58
##  5 EJE CENTRAL LAZARO CARDENAS    53
##  6 INSURGENTES SUR                42
##  7 AV. INSURGENTES SUR            40
##  8 PERIFERICO SUR                 40
##  9 CALZADA ERMITA IZTAPALAPA      36
## 10 ERMITA IZTAPALAPA              34
## # ... with 9,640 more rows
```

```
# Con la funcion arrange y la funcion desc() de ordenamiento descendiente.
carpetas %>%
  count(calle_hechos) %>%
  arrange(desc(n))
```

```
## # A tibble: 9,650 x 2
##    calle_hechos                   n
##    <chr>                      <int>
##  1 CALZADA DE TLALPAN            65
##  2 CALZADA DE GUADALUPE          63
##  3 AVENIDA TLAHUAC               62
##  4 CALZADA IGNACIO ZARAGOZA      58
##  5 EJE CENTRAL LAZARO CARDENAS   53
##  6 INSURGENTES SUR               42
##  7 AV. INSURGENTES SUR           40
##  8 PERIFERICO SUR                40
##  9 CALZADA ERMITA IZTAPALAPA     36
## 10 ERMITA IZTAPALAPA             34
## # ... with 9,640 more rows
```

```
# Guardando la nueva base en un objeto llamado noCaminesPorAhi, pasando el argumento -n a la funcion ar
noCaminesPorAhi <- carpetas %>%
  count(calle_hechos) %>%
  arrange(-n)
noCaminesPorAhi
```

```
## # A tibble: 9,650 x 2
##    calle_hechos                   n
##    <chr>                      <int>
##  1 CALZADA DE TLALPAN            65
##  2 CALZADA DE GUADALUPE          63
##  3 AVENIDA TLAHUAC               62
##  4 CALZADA IGNACIO ZARAGOZA      58
##  5 EJE CENTRAL LAZARO CARDENAS   53
##  6 INSURGENTES SUR               42
##  7 AV. INSURGENTES SUR           40
##  8 PERIFERICO SUR                40
##  9 CALZADA ERMITA IZTAPALAPA     36
## 10 ERMITA IZTAPALAPA             34
## # ... with 9,640 more rows
```

```
# Ahora, a esta base le recortamos las 5 calles más peligrosas y le renombramos las variables.

noCaminesPorAhi %>%
  head(n = 5) %>%
```

```
  rename(Calles = calle_hechos,
         No_Delitos = n) # Primero nombre nuevo, luego igual, y luego nombre viejo
```

```
## # A tibble: 5 x 2
##   Calles                      No_Delitos
##   <chr>                            <int>
## 1 CALZADA DE TLALPAN                  65
## 2 CALZADA DE GUADALUPE                63
## 3 AVENIDA TLAHUAC                     62
## 4 CALZADA IGNACIO ZARAGOZA            58
## 5 EJE CENTRAL LAZARO CARDENAS         53
```

Utilizar los verbos select y filter

```
# Seleccionamos la columna fecha_hechos y todas las variables que contengan el caracter 2.
carpetas %>%
  select(fecha_hechos, contains("2"))
```

```
## # A tibble: 16,623 x 2
##    fecha_hechos        calle_hechos2
##    <dttm>              <chr>
##  1 2019-08-03 17:40:00 <NA>
##  2 2019-08-03 15:42:00 SAINT SAENZ
##  3 2019-08-02 11:00:00 DESIDERIO PEÑA
##  4 2019-08-02 20:25:00 ANILLO PERIFERICO (BLVD. ADOLFO RUIZ CORTINES)
##  5 2019-08-02 17:40:00 <NA>
##  6 2019-08-03 15:30:00 <NA>
##  7 2019-08-03 19:00:00 <NA>
##  8 2019-08-02 06:00:00 <NA>
##  9 2019-08-03 15:30:00 <NA>
## 10 2019-08-02 07:00:00 EJIDO SAN ANTONIO TOMATLAN
## # ... with 16,613 more rows
```

```
# Seleccionamos la columna fecha_hechos y todas las columnas que terminen con ito, las que empiecen con
baseNueva <- carpetas %>%
  select(fecha_hechos,
         ends_with("ito"),
         starts_with("a"),
         contains("fec"))

# A esta base nueva le filtramos lo siguiente:
# 1. Las observaciones que sean de la alcaldia Benito Juarez o de la Alcaldia Coyoacan,
# 2. de estas, las que correspondan solamente a la categoria de Delitos de Bajo Impacto
# 3. y de estas, las que hayan ocurrido entre el 2 y el 30 de agosto

baseNueva %>%
  filter(alcaldia_hechos == "BENITO JUAREZ" | alcaldia_hechos == "COYOACAN",
         categoria_delito == "DELITO DE BAJO IMPACTO",
         fecha_hechos > "2019-08-01" & fecha_hechos < "2019-08-31 00:00:00")
```

```
## # A tibble: 1,920 x 8
##    fecha_hechos        delito categoria_delito ao_hechos agencia
##    <dttm>              <chr>  <chr>                <dbl> <chr>
##  1 2019-08-03 15:30:00 AMENA~ DELITO DE BAJO ~      2019 BJ-5
```

```
##  2 2019-08-02 07:00:00 ROBO ~ DELITO DE BAJO ~      2019 COY-3
##  3 2019-08-02 18:00:00 ROBO ~ DELITO DE BAJO ~      2019 TUR-2
##  4 2019-08-03 18:30:00 ROBO ~ DELITO DE BAJO ~      2019 COY-3
##  5 2019-08-01 16:00:00 ROBO ~ DELITO DE BAJO ~      2019 COY-2
##  6 2019-08-03 18:00:00 ABUSO~ DELITO DE BAJO ~      2019 FDS-5
##  7 2019-08-04 06:30:00 DAÑO ~ DELITO DE BAJO ~      2019 COY-1
##  8 2019-08-04 04:00:00 ROBO ~ DELITO DE BAJO ~      2019 COY-2
##  9 2019-08-04 09:50:00 ROBO ~ DELITO DE BAJO ~      2019 COY-2
## 10 2019-08-02 21:39:00 ROBO ~ DELITO DE BAJO ~      2019 BJ-1
## # ... with 1,910 more rows, and 3 more variables: alcaldia_hechos <chr>,
## #   ao_inicio <dbl>, fecha_inicio <dttm>
```

```r
# Estos dos cachitos de codigo se pueden escribir en una sola cadena:
nuevaNuevaBase <- carpetas %>%
  select(fecha_hechos,
         ends_with("ito"),
         starts_with("a"),
         contains("fec")) %>%
  filter(alcaldia_hechos == "BENITO JUAREZ" | alcaldia_hechos == "COYOACAN",
         categoria_delito == "DELITO DE BAJO IMPACTO",
         fecha_hechos > "2019-08-01" & fecha_hechos < "2019-08-31 00:00:00")

# Todo Junto.
```

Suerte en el exámen chicos.