Estudi dels accidents de trànsit als Estats Units d'Amèrica

Visualització de dades - Pràctica 2

Héctor Gutiérrez Muñoz

9 de juny, 2021

Sumari

1 Càrrega del fitxer

En primer lloc, carreguem els fitxers CSV. El dataset principal es pot trobar a https://www.kaggle.com/sobhanmoosavi/us-accidents. L'altre, de la població als Estats Units està disponible a https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/totals/nst-est2019-01.xlsx.

```
accidents_data <- read.csv("US_Accidents_Dec20_Updated.csv")

population_data <- data.frame(read_excel(
  path = "nst-est2019-01.xlsx",
  range = c("A10:M60"),
  col_names = FALSE))
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...
```

```
# Només ens interessen dues columnes d'aquest dataset

population_data <- population_data[,c(1,13)]
colnames(population_data) <- c("state", "population")
```

2 Primera exploració

Veiem uns exemples del primer dataset:

```
kable(
  t(head(accidents_data,2)),
  format = 'latex',
  booktabs = TRUE
) %>%
 kableExtra::kable_styling(latex_options = c("scale_down","hold_position"))
```

També del segon:

```
kable(
  head(population_data),
  format = 'latex',
  booktabs = TRUE
) %>%
 kableExtra::kable_styling(latex_options = c("hold_position"))
```

3 Preprocessament

L'únic preprocessament que cal fer en aquest cas és fer el join dels dos datsets pel estat. Per fer-lo, primer s'han de posar les dues variables en el mateix format.

```
# Traiem el punt que hi ha a l'inici dels noms

population_data$state_trim <- substring(population_data$state, 2)

# Traiem la dada del Districte de Columbia, que no és a les variables state.abb
# i state.name de R

pop_DC <- population_data[population_data$state_trim == "District of Columbia", 2]

population_data <-
  population_data[-which(population_data$state_trim == "District of Columbia"),]

# Fem la conversió a abreviatures

population_data$state_abb <- state.abb[which(state.name == population_data$state_trim)]

# Tornem a introduir la dada del Districte de Columbia

population_data <- rbind(population_data, c("", pop_DC, "District of Columbia", "DC"))
```

Un cop fet això, podem fer el join:

```
merged_df <- merge(accidents_data,
                   population_data,
                   by.x = "State",
```

	1	2
ID	A-1	A-2
Severity	2	2
Start_Time	2019-05-21 08:29:55	2019-10-07 17:43:09
End_Time	2019-05-21 09:29:40	2019-10-07 19:42:50
Start_Lat	34.80887	35.09008
Start_Lng	-82.26916	-80.74556
End_Lat	34.80887	35.09008
End_Lng	-82.26916	-80.74556
Distance.mi.	0	0
Description	Accident on Tanner Rd at Pennbrooke Ln.	Accident on Houston Branch Rd at Providence Branch Ln.
Number	439	3299
Street	Tanner Rd	Providence Branch Ln
Side	R	R
City	Greenville	Charlotte
County	Greenville	Mecklenburg
State	SC	NC
Zipcode	29607-6027	28270-8560
Country	US	US
Timezone	US/Eastern	US/Eastern
Airport_Code	KGMU	KEQY
Weather_Timestamp	2019-05-21 08:53:00	2019-10-07 17:53:00
Temperature.F.	76	76
Wind_Chill.F.	76	76
Humidity...	52	62
Pressure.in.	28.91	29.30
Visibility.mi.	10	10
Wind_Direction	N	VAR
Wind_Speed.mph.	7	3
Precipitation.in.	0	0
Weather_Condition	Fair	Cloudy
Amenity	False	False
Bump	False	False
Crossing	False	False
Give_Way	False	False
Junction	False	False
No_Exit	False	False
Railway	False	False
Roundabout	False	False
Station	False	False
Stop	False	False
Traffic_Calming	False	False
Traffic_Signal	False	False
Turning_Loop	False	False
Sunrise_Sunset	Day	Day
Civil_Twilight	Day	Day
Nautical_Twilight	Day	Day
Astronomical_Twilight	Day	Day

state	population
.Alabama	4903185
.Alaska	731545
.Arizona	7278717
.Arkansas	3017804
.California	39512223
.Colorado	5758736

```
                  by.y = "state_abb")

head(merged_df)
```

```
##   State        ID Severity          Start_Time            End_Time Start_Lat
## 1    AL  A-639502        4 2017-09-22 21:46:42 2017-09-23 03:46:42  33.68182
## 2    AL A-2162021        3 2019-04-12 19:27:04 2019-04-12 20:56:46  32.67388
## 3    AL A-2813019        2 2017-12-11 07:59:43 2017-12-11 08:29:30  33.51469
## 4    AL  A-855293        2 2020-11-13 01:54:00 2020-11-13 04:15:00  32.46024
## 5    AL A-1889735        2 2017-09-15 18:42:52 2017-09-15 19:12:27  30.49452
## 6    AL A-1018580        2 2019-05-02 07:53:04 2019-05-02 09:07:54  30.68940
##   Start_Lng  End_Lat   End_Lng Distance.mi.
## 1 -87.09209 33.67638 -87.07980        0.800
## 2 -85.33090 32.67388 -85.33090        0.000
## 3 -86.78768 33.51469 -86.78768        0.000
## 4 -86.38945 32.46023 -86.38899        0.027
## 5 -88.21700 30.49452 -88.21700        0.000
## 6 -88.11458 30.68940 -88.11458        0.000
##                                                                Description
## 1         Closed at CR-81/Sharon Blvd/Exit 78 - Road closed due to accident.
## 2                             Accident on I-85 Southbound at Exit 64 US-29.
## 3                                       Accident on 5th Ave at 30th St.
## 4  Incident on COBBS FORD RD EB near I-65 Right lane blocked. Expect delays.
## 5 Accident on County Hwy-23 Padgett Switch Rd at County Hwy-24 Half Mile Rd.
## 6                                    Accident on Dauphin St at Sage Ave.
##   Number        Street Side       City    County    Zipcode Country   Timezone
## 1    NaN        AL-4 E    R    Quinton    Walker      35130      US US/Central
## 2    NaN        I-85 N    R    Opelika       Lee      36801      US US/Central
## 3   2999     5th Ave S    R Birmingham Jefferson 35233-2916      US US/Central
## 4    NaN Cobbs Ford Rd    R Prattville    Elmore      36066      US US/Central
## 5    NaN   McDonald Rd    R  Irvington    Mobile      36544      US US/Central
## 6    NaN    S Sage Ave    R     Mobile    Mobile      36606      US US/Central
##   Airport_Code   Weather_Timestamp Temperature.F. Wind_Chill.F. Humidity...
## 1         KJFX 2017-09-22 21:55:00           71.6           NaN         100
## 2         KAUO 2019-04-12 19:56:00           69.0            69          73
## 3         KBHM 2017-12-11 07:53:00           37.0           NaN          82
## 4         KMXF 2020-11-13 01:56:00           59.0            59         NaN
## 5         KMOB 2017-09-15 18:56:00           79.0           NaN          79
## 6         KBFM 2019-05-02 07:53:00           73.0            73          84
##   Pressure.in. Visibility.mi. Wind_Direction Wind_Speed.mph. Precipitation.in.
## 1        30.00              3           Calm             NaN               NaN
## 2        29.18             10           CALM             0.0                 0
## 3        30.24             10           Calm             NaN               NaN
## 4        29.87              9            NNW             7.0                 0
## 5        29.97             10             SE             4.6               NaN
## 6        30.09             10             NE             3.0                 0
##   Weather_Condition Amenity  Bump Crossing Give_Way Junction No_Exit Railway
## 1             Clear   False False    False    False    False   False   False
## 2              Fair   False False    False    False    False   False   False
## 3             Clear   False False    False    False    False   False   False
## 4            Cloudy   False False    False    False    False   False   False
## 5  Scattered Clouds   False False    False    False    False   False   False
## 6              Fair   False False    False    False    False   False   False
```

```
##   Roundabout Station  Stop Traffic_Calming Traffic_Signal Turning_Loop
## 1      False   False False           False          False        False
## 2      False   False False           False          False        False
## 3      False   False False           False          False        False
## 4      False   False False           False          False        False
## 5      False   False False           False           True        False
## 6      False   False False           False           True        False
##   Sunrise_Sunset Civil_Twilight Nautical_Twilight Astronomical_Twilight
## 1          Night          Night             Night                 Night
## 2          Night            Day               Day                   Day
## 3            Day            Day               Day                   Day
## 4          Night          Night             Night                 Night
## 5            Day            Day               Day                   Day
## 6            Day            Day               Day                   Day
##      state population state_trim
## 1 .Alabama    4903185    Alabama
## 2 .Alabama    4903185    Alabama
## 3 .Alabama    4903185    Alabama
## 4 .Alabama    4903185    Alabama
## 5 .Alabama    4903185    Alabama
## 6 .Alabama    4903185    Alabama
```

Es comprova que el join s'ha fet correctament i que el nombre de registres és el mateix que al dataset original:

```
dim(accidents_data)[1] == dim(merged_df)[1]
```

```
## [1] TRUE
```

Com que el dataset és força gran, en treurem les columnes que no es faran servir a la visualització per intentar millorar la performance.

```
merged_df <- merged_df[, c(
  "Severity", "Start_Time", "End_Time", "Start_Lat", "Start_Lng",
  "Distance.mi.", "State", "Temperature.F.", "Wind_Chill.F.", "Humidity...",
  "Pressure.in.", "Visibility.mi.", "Wind_Speed.mph.", "Precipitation.in.",
  "Weather_Condition", "Amenity", "Bump", "Crossing", "Give_Way", "Junction",
  "No_Exit", "Railway", "Roundabout", "Station", "Stop", "Traffic_Calming",
  "Traffic_Signal", "Turning_Loop", "Sunrise_Sunset", "population", "state_trim"
)]
```

Les dades de l'any 2016 no són força acurades, hi ha molts menys accidents que la resta d'anys. Per no desvirtuar l'anàlisi, traiem aquests registres del dataset.

```
merged_df <- merged_df[substring(merged_df$Start_Time,1,4) != "2016", ]
```

4 Fitxer de sortida

Finalment, les dades preprocessades es desaran a un fitxer.

```r
write.csv2(merged_df,
           "accidents_final.csv",
           row.names = FALSE,
           na = "")
```