

Herramientas de procesamiento de datos a utilizar

Rstudio proporciona diferentes herramientas para procesar los datos, como es el caso de janitor.

Ventajas:

`tabyl()`: entrega el número y porcentaje de apariciones de un mismo valor en una columna.

`levels()`: entrega los datos almacenados según su valor, al igual que `tabyl()`, pero sin su número de repeticiones ni porcentaje

`replace()`: modificar información del dataset

`subset()`: elimina o guarda datos con restricciones deseadas

Rstudio permite visualizar el dataset en una ventana de tipo excel de forma completa

Desventajas:

`tabyl()`: no permite conocer el id correspondiente a cada caso.

`levels()`: no nos deja conocer información de identificadores de posición de las consultas.

`replace()`: permite modificar uno por uno y no un grupo

Rstudio nos permite ver sólo los primeros 1000 elementos de nuestras consultas, además, al visualizar el dataset completo en una ventana no podemos editarlo directamente.

Dataset

Instancias: 77946

Atributos: 28

Tipo de datos: data.frame

numeric

factor

valores perdidos: datos con locación vacía

valores con ruido: datos con caracteres no correspondientes en tweet, estado o locación

outliers: datos con ubicación correspondiente a una locación fuera de Estados Unidos.

Limpieza de datos

Carga de datos en r:

```
datos <- read.csv(file = "train.csv",header = TRUE, sep = ",")
```

Visualización del dataset:

datos x																											
	id	tweet	state	location	s1	s2	s3	s4	s5	w1	w2	w3	w4	k1	k2	k3	k4	k5	k6	k7							
1	1	Jazz for a Rainy Afternoon: (link)	oklahoma	Oklahoma	0.000	0.000	1.000	0.000	0.000	0.800	0.000	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00							
2	2	RT: @mention: I love rainy days.	florida	Miami-Ft. Lauderdale	0.000	0.000	0.000	1.000	0.000	0.196	0.000	0.804	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00							
3	3	Good Morning Chicago! Time to kick the Windy City in t...	idaho		0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	1.00							
4	6	Preach lol! ; RT @mention: #aliwantis this type of weat...	minnesota	Minneapolis-St. Paul	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.60							
5	9	@mention good morning sunshine	rhode island	Purgatory	0.000	0.000	0.000	0.403	0.597	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00							
6	10	Rather be storm chasing.	district of columbia	Washington, DC	0.000	0.000	0.600	0.000	0.400	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00	0.00							
7	11	#WEATHER: 1:54 pm : 61.0F, Feels 60F, 29.98% Humidity: ...	michigan	St Louis, MI, USA	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.203	0.176	0.000	0.376	0.00	0.42						
8	13	Father. Son. Holy ghost. Amen... hope I come out alive =...	connecticut	Hartford & New Haven	0.000	0.193	0.000	0.000	0.807	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.00	0.00						
9	14	Mild Monday best weather day of the week in Oklahom...	oklahoma	Oklahoma City	0.000	0.000	0.000	1.000	0.000	0.599	0.000	0.401	0.000	0.000	0.000	0.000	0.000	0.194	0.000	0.60							

Visualización alternativa:

```
> head(datos)
  id
1  1
2  2
3  3
4  6
5  9
6 10
```

```

                                tweet
1      Jazz for a Rainy Afternoon: {link}
2      RT: @mention: I love rainy days.
3      Good Morning Chicago! Time to kick the win
dy City in the nuts and head back west!
4 Preach lol! :) RT @mention: #alliwantis this type of weather all the time.. I live for
beautiful days like this! #minneapolis
5      @mention good morning sunshine
6      Rather be storm chasing.

                                state      location s1 s2 s3 s4 s5 w1 w2 w3 w4
1      oklahoma      oklahoma 0 0 1.0 0.000 0.000 0.800 0 0.200 0
2      florida      Miami-Ft. Lauderdale 0 0 0.0 1.000 0.000 0.196 0 0.804 0
3      idaho      0 0 0.0 0.000 1.000 0.000 0 1.000 0
4      minnesota      Minneapolis-St. Paul 0 0 0.0 1.000 0.000 1.000 0 0.000 0
5      rhode island      Purgatory 0 0 0.0 0.403 0.597 1.000 0 0.000 0
6 district of columbia      washington, DC 0 0 0.6 0.000 0.400 0.000 0 1.000 0
k1 k2 k3 k4 k5 k6 k7 k8 k9 k10 k11 k12 k13 k14 k15
1 0 0 0 0 0 0 0 0.000 0 0.000 1 0 0 0.000 0 0
2 0 0 0 0 0 0 0 0.000 0 0.000 1 0 0 0.000 0 0
3 0 0 0 0 0 0 0 1.000 0 0.000 0 0 0 0.000 0 0
4 0 0 0 0 0 0 0 0.604 0 0.196 0 0 0 0.201 0 0
5 0 0 0 0 0 0 0 0.000 0 0.000 0 0 0 1.000 0 0
6 0 0 0 0 0 0 0 0.000 0 0.000 0 0 1 0.000 0 0

```

Uso de tabyl():

```

> tabyl(datos$location)
                                datos$location      n      percent
-- bull cityyyy !      2 2.565879e-05
--Unknown--      1 1.282939e-05
- Columbus (:      1 1.282939e-05
- jersey city [:      1 1.282939e-05
-Behind U wit Sum Shades on-      2 2.565879e-05
-Houston , TX      1 1.282939e-05
-illinois-chicago-      1 1.282939e-05
-Phoenix; AZ      1 1.282939e-05
[chicago] â~œâ~jâ~ž      1 1.282939e-05
[wv]      2 2.565879e-05
4, MS      1 1.282939e-05
72401      1 1.282939e-05

> tabyl(datos$state)
                                datos$state      n      percent
alabama 1577 0.020231955
alaska 714 0.009160188
arizona 1575 0.020206297
arkansas 808 0.010366151
california 3765 0.048302671
colorado 1712 0.021963924
connecticut 1548 0.019859903
delaware 749 0.009609217
district of columbia 782 0.010032587
florida 3659 0.046942755
georgia 1519 0.019487851
hawaii 765 0.009814487
idaho 752 0.009647705

```

Con esto podemos concluir que existen locaciones vacías que representan el 14% del dataset, pero no estados vacíos.

```
> head(levels(datos$location))
[1] ""                                "-- bull cityyy !"
[3] "--unknown--"                   "- columbus (:)"
[5] "- jersey city (:)"             "-Behind U wit Sum shades On-"
```

```
datos2 <- subset(datos, location != '')
```

Data	
datos	77946 obs. of 28 variables
datos2	67018 obs. of 28 variables

Inicial

Alabama	75	1.119102e-03
ALABAMA	4	5.968546e-05
Alabama	2	2.984273e-05
Alabama State University	1	1.492136e-05
Alabama, Montgomery	2	2.984273e-05
Alabama, Tuscaloosa	1	1.492136e-05
Alabama, USA	1	1.492136e-05
Alabama...=)	1	1.492136e-05
Alabaster, AL	4	5.968546e-05
Alamogordo, NM	3	4.476409e-05

```
datos2$location[datos2$location=="Alabama...")<- "Alabama"
```

Final

Alabama	76	1.134024e-03	1.134058e-03
ALABAMA	4	5.968546e-05	5.968724e-05
Alabama	2	2.984273e-05	2.984362e-05
Alabama State University	1	1.492136e-05	1.492181e-05
Alabama, Montgomery	2	2.984273e-05	2.984362e-05
Alabama, Tuscaloosa	1	1.492136e-05	1.492181e-05
Alabama, USA	1	1.492136e-05	1.492181e-05
Alabama...=)	0	0.000000e+00	0.000000e+00
Alabaster, AL	4	5.968546e-05	5.968724e-05
Alamogordo, NM	3	4.476409e-05	4.476543e-05