Wikia census: cleaning and analysis

Abel Serrano Juste

23 de diciembre, 2018

Índice general

Introducción **1**

Problema a resolver **1**

Fuentes de datos **2**

Descripción del Dataset **2**
Wikia census dataset 2
Wikia page views dataset 7

Limpieza **8**

Transformación **10**

Integración **11**

Resultados **13**

Referencias **13**

Introducción

En esta práctica, vamos a realizar un proyecto analítico de ciencia de datos sobre el ecosistema de las wikis.

Según la definición de Wikipedia para wiki:

> El término wiki (proviene del hawaiano wiki, «rápido») alude al nombre que recibe una comunidad virtual, cuyas páginas son editadas directamente desde el navegador, donde los mismos usuarios crean, modifican, corrigen o eliminan contenidos que, generalmente, comparten.

Las wikis son un interesante objeto de estudio puesto que permiten investigar la colaboración masiva de usuarios online para crear un contenido común.

Utilizaremos los términos usuario y editor indistantemente, puesto que en el contexto de una wiki se pueden entender como sinónimos.

Problema a resolver

Analizar la actividad y diversidad de las wikis alojadas en el servicio Wikia.

Fuentes de datos

Para esta práctica, vamos a usar dos datasets: **Wikia census** y **Wikia page views**; ambos disponibles en mi cuenta de kaggle.

1. El censo de Wikia. Se trata de un dataset de un conjunto de 300k wikis que corresponde a todas las wikis alojadas en Wikia. Este dataset contiene datos descriptivos de cada wiki como: número de páginas, número de usuarios, número de ediciones, etc. Los métodos de extracción y la información proporcionada en este censo está explicada en el paper: A Wikia census: motives, tools and insights (Jimenez-Diaz, Serrano, y Arroyo 2018).
2. Wikia page views. Se trata de una captura de datos, realizada mediante web scrapping de todas las wikis de Wikia, que contiene el número de visitas para cada una de las wikis de Wikia en las últimas cuatro semanas. Este dataset se obtuvo para la práctica anterior de esta misma asignatura y el código fuente para su obtención está en el este repositorio de Github: https://github.com/Akronix/scrap_wikia_page_views.

Descripción del Dataset

Como hemos explicado previamente, vamos a usar dos datasets: el censo de wikia y los números de páginas visitadas.

Wikia census dataset

El primer paso consistirá en cargar los datos:

```
# Cargamos el juego de datos
wikis<-read.csv("data/20181019-wikia_stats_users_birthdate.csv",header=T,sep=",")
```

A continuación, haremos una breve descripción de los datos, ya que nos interesa tener una idea general de los datos que disponemos. Para ello, primero calcularemos las dimensiones de nuestra base de datos y mostraremos una muestra de los datos para interpretar qué tipos de atributos tenemos.

```
dim(wikis)
```

```
## [1] 277795     32
```

Disponemos de datos de 277795 wikis (filas) con 32 atributos sobre cada uno de ellos (columnas).

```
head(wikis)
```

```
##                                                 url
## 1                           http://spellmagotm.wikia.com
## 2 http://2017-monster-energy-nascar-cup-series.wikia.com
## 3                     http://10low46japreligion.wikia.com
## 4                        http://indigo-showdown.wikia.com
## 5                             http://animewiki2.wikia.com
## 6                   http://elena-ofavalor-fans.wikia.com
##         creation_date                                          domain
## 1 2012-05-01 13:58:13                           spellmagotm.wikia.com
## 2 2017-07-24 22:35:31 2017-monster-energy-nascar-cup-series.wikia.com
## 3 2009-09-15 23:21:34                     10low46japreligion.wikia.com
## 4 2014-05-30 15:43:23                        indigo-showdown.wikia.com
## 5 2011-02-18 23:15:14                             animewiki2.wikia.com
## 6 2018-09-12 17:52:40                   elena-ofavalor-fans.wikia.com
```

```
##   founding_user_id headline       hub      id lang language
## 1          5069110              Games  529058   en       en
## 2         32529801                 TV 1601247   en       en
## 3          1602876          Lifestyle   52061   en       en
## 4         25001469              Games  982346   en       en
## 5          1160460                 TV  221590   en       en
## 6         36888175                 TV 1806604   en       en
##                                                 name stats.activeUsers
## 1                                  Spellmagotm Wiki                 0
## 2            2017 Monster Energy NASCAR Cup Series Wiki              0
## 3 Ancient Japanese Religion (Daramalan Assignment) Wiki              0
## 4                                  Indigo showdown Wiki              0
## 5                                       Animewiki2 Wiki              2
## 6                                Elena ofavalor fans Wiki            -1
##   stats.admins stats.articles stats.discussions stats.edits stats.images
## 1            1              8                 0         275           11
## 2            1             45                 0         230           40
## 3            1             71                 1         470           20
## 4            1             17                 1         242           11
## 5            1            102                NA        6365         4938
## 6            1              1                 0          85            0
##   stats.pages stats.users stats.videos
## 1         229    16246020            0
## 2         153    16555972            0
## 3         310    16117576            4
## 4         156    16387583            2
## 5        5305    15397437            2
## 6          83    15565473            0
##                                                 title         topic
## 1                                  Spellmagotm Wiki        Gaming
## 2            2017 Monster Energy NASCAR Cup Series Wiki Entertainment
## 3 Ancient Japanese Religion (Daramalan Assignment) Wiki    Philosophy
## 4                                  Indigo showdown Wiki        Gaming
## 5                                       Animewiki2 Wiki         Anime
## 6                                Elena ofavalor fans Wiki Entertainment
##   wam_score stats.nonarticles users_1 users_5 users_10 users_20 users_50
## 1    0.0000               221       5       1        1        1        1
## 2    0.0000               108       6       3        1        1        1
## 3    0.0000               239       7       2        1        1        1
## 4    0.0000               139       9       4        3        3        1
## 5    0.0161              5203      33      21       16       14        9
## 6    0.0000                82       3       0        0        0        0
##   users_100 bots               birthDate  datetime.birthDate
## 1         0    5       14:14, May 1, 2012 2012-05-01 14:14:00
## 2         1    4     22:35, July 24, 2017 2017-07-24 22:35:00
## 3         1    5 23:21, September 15, 2009 2009-09-15 23:21:00
## 4         0    5      15:43, May 30, 2014 2014-05-30 15:43:00
## 5         6    3  23:15, February 18, 2011 2011-02-18 23:15:00
## 6         0    2 17:52, September 12, 2018 2018-09-12 17:52:00
```

```
str(wikis)
```

```
## 'data.frame':    277795 obs. of  32 variables:
## $ url          : Factor w/ 277795 levels "http://0002oifos.wikia.com",..: 225470 532 147 126585 9385 594
## $ creation_date    : Factor w/ 276956 levels "2001-01-15 00:00:00",..: 53387 183565 10462 107722 30199 27
```

```
## $ domain        : Factor w/ 277794 levels "0002oifos.wikia.com",..: 225469 532 147 126584 9385 59413 232
##  $ founding_user_id  : num  5069110 32529801 1602876 25001469 1160460 ...
## $ headline        : Factor w/ 8152 levels "","              .",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ hub               : Factor w/ 8 levels "Books","Comics",..: 3 8 4 3 8 8 4 8 8 3 ...
##  $ id                : num  529058 1601247 52061 982346 221590 ...
## $ lang          : Factor w/ 79 levels "aa","af","am",..: 20 20 20 20 20 20 20 20 20 20 ...
## $ language       : Factor w/ 79 levels "aa","af","am",..: 20 20 20 20 20 20 20 20 20 20 ...
## $ name          : Factor w/ 258804 levels "   ","  ",..: 175995 608 8765 90942 10708 56819 174447 217529 1
##  $ stats.activeUsers : num  0 0 0 0 2 -1 0 0 0 0 ...
##  $ stats.admins      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ stats.articles    : num  8 45 71 17 102 1 6 7 6 4 ...
##  $ stats.discussions : num  0 0 1 1 NA 0 1 0 0 0 ...
##  $ stats.edits       : num  275 230 470 242 6365 ...
##  $ stats.images      : num  11 40 20 11 4938 ...
##  $ stats.pages       : num  229 153 310 156 5305 ...
##  $ stats.users       : num  16246020 16555972 16117576 16387583 15397437 ...
##  $ stats.videos      : num  0 0 4 2 2 0 0 0 0 0 ...
## $ title         : Factor w/ 258804 levels "   ","  ",..: 175995 608 8765 90942 10708 56819 174447 217529 1
##  $ topic           : Factor w/ 29 levels "","Anime","Auto",..: 12 8 19 12 2 8 16 9 9 1 ...
##  $ wam_score         : num  0 0 0 0 0.0161 0 0 0 0 0 ...
##  $ stats.nonarticles : num  221 108 239 139 5203 ...
##  $ users_1           : int  5 6 7 9 33 3 9 7 4 4 ...
##  $ users_5           : int  1 3 2 4 21 0 1 2 0 1 ...
##  $ users_10          : int  1 1 1 3 16 0 0 2 0 1 ...
##  $ users_20          : int  1 1 1 3 14 0 0 2 0 0 ...
##  $ users_50          : int  1 1 1 1 9 0 0 0 0 0 ...
##  $ users_100         : int  0 1 1 0 6 0 0 0 0 0 ...
##  $ bots              : int  5 4 5 5 3 2 5 6 2 2 ...
## $ birthDate      : Factor w/ 267762 levels "00:00 10 mar 2013",..: 123272 240694 251221 141386 249826 170
## $ datetime.birthDate: Factor w/ 261494 levels "0001-07-28 20:25:00",..: 51943 178041 10093 104783 29390 1
```

En base a la muestra, al conocimiento sobre el campo en el que estamos trabajando (wikis) y a la descripción proporcionada en el paper *"A Wikia census: motives, tools and insights"*, deducimos los siguientes atributos:

- url: url de la wiki
- creation_date: fecha de creación de la wiki en un timestamp
- domain: dominio web de la wiki
- founding_user_id: user id del fundador de la wiki
- headline: ??
- hub: Categoría de la wiki dentro de las definidas por Wikia.
- id: id de la wiki
- lang y language: idioma de la wiki
- name: Nombre propio de la wiki
- stats.activeUsers: número de usuarios activos en el último mes. Los usuarios activos son los usuarios que han hecho al menos una acción (una edición) en los últimos 30 días.
- stats.admins: número de usuarios administradores.
- stats.articles: número de artículos de la wiki.
- stats.discussions: ???
- stats.edits: número de ediciones en la wiki.
- stats.images: número de imágenes subidas.
- stats.pages: número de páginas de la wiki.
- stats.users: número de usuarios registrados en toda Wikia (potencialmente, cualquier podría user usuario de cada wiki porque los usuarios se registran a nivel de toda Wikia).
- stats.videos: número de videos subidos.
- title: título de la wiki

- topic: Temática de la wiki definida por el administrador de la wiki.
- wam_score: Puntuación que le da Wikia a las wikis: http://community.wikia.com/wiki/WAM/FAQ
- stats.nonarticles: número de páginas no artículos en la wiki.
- users_{1,5,10,20,50,100}: Número de usuarios con al menos {una, cinco, diez, veinte, cincuenta, cien} edición(es).
- bots: Número de usuarios de tipo bot (no humanos)
- birthdate: fecha de creación de la wiki en formato natural
- datetime.birthDate: fecha de creación de la wiki en formato datetime de Python

Para terminar con el estudio previo de los datos, pedimos a R que nos muestre un resumen de cómo están distribuidos los valores de los atributos:

```
summary(wikis)
```

```
##                                 url                      creation_date
##  http://0002oifos.wikia.com       :     1   2006-03-16 13:42:45:   388
##  http://001-game-creator.wikia.com:     1   2004-11-11 23:33:14:    79
##  http://001littlebighelp.wikia.com:     1   2006-10-28 11:16:22:    14
##  http://007fanon.wikia.com        :     1   2017-10-16 21:02:01:    12
##  http://007goldeneye.wikia.com    :     1   2017-10-02 21:37:50:     8
##  http://007-james-bond.wikia.com  :     1   2018-02-08 13:53:15:     7
##  (Other)                          :277789   (Other)            :277287
##                          domain       founding_user_id
##  cliff-side.wikia.com      :     2   Min.   :       0
##  0002oifos.wikia.com       :     1   1st Qu.: 4948217
##  001-game-creator.wikia.com:     1   Median :25524286
##  001littlebighelp.wikia.com:     1   Mean   :20302830
##  007fanon.wikia.com        :     1   3rd Qu.:32701075
##  007goldeneye.wikia.com    :     1   Max.   :36902116
##  (Other)                   :277788   NA's   :35
##                    headline             hub                id
##                        :269610   Games    :104758   Min.   :      1
##  Assassin's Creed Wiki :     3   Lifestyle: 65294   1st Qu.: 659216
##  The Elder Scrolls Wiki:     3   TV       : 44343   Median :1278192
##  Alice Wiki            :     2   Books    : 22619   Mean   :1137079
##  Animal Crossing Wiki  :     2   Comics   : 14545   3rd Qu.:1685929
##  Bakugan Wiki          :     2   Movies   : 14174   Max.   :1807291
##  (Other)               :  8173   (Other)  : 12062
##       lang           language                        name
##  en     :198927   en     :198927   Test Wiki          :    67
##  es     : 30372   es     : 30372   Naruto Wiki        :    59
##  ru     : 12967   ru     : 12967   Harry Potter Wiki  :    54
##  de     :  9176   de     :  9176   Pokemon Wiki       :    48
##  pl     :  7582   pl     :  7582   Minecraft Wiki     :    44
##  fr     :  7029   fr     :  7029   Final Frontier Wiki:    39
##  (Other): 11742   (Other): 11742   (Other)            :277484
##  stats.activeUsers   stats.admins     stats.articles
##  Min.   :  -1.000   Min.   :  0.000   Min.   :      0.0
##  1st Qu.:   0.000   1st Qu.:  1.000   1st Qu.:      4.0
##  Median :   0.000   Median :  1.000   Median :     10.0
##  Mean   :   0.823   Mean   :  1.481   Mean   :    122.1
##  3rd Qu.:   0.000   3rd Qu.:  1.000   3rd Qu.:     30.0
##  Max.   :7311.000   Max.   :248.000   Max.   :2483023.0
##
##  stats.discussions   stats.edits        stats.images
```

```
##  Min.   :    0.00   Min.   :       0   Min.   :   -15.0
##  1st Qu.:    0.00   1st Qu.:     116   1st Qu.:     1.0
##  Median :    0.00   Median :     245   Median :    10.0
##  Mean   :    2.97   Mean   :    2755   Mean   :   210.7
##  3rd Qu.:    1.00   3rd Qu.:     470   3rd Qu.:    40.0
##  Max.   :77051.00   Max.   :23302062   Max.   :784356.0
##  NA's   :52295
##   stats.pages       stats.users        stats.videos
##  Min.   :     -1   Min.   :      -1   Min.   :    0.00
##  1st Qu.:    101   1st Qu.:16140740   1st Qu.:    0.00
##  Median :    172   Median :16374841   Median :    0.00
##  Mean   :   1046   Mean   :16229539   Mean   :    7.45
##  3rd Qu.:    285   3rd Qu.:16555972   3rd Qu.:    0.00
##  Max.   :7727191   Max.   :23922667   Max.   :51799.00
##
##                  title                   topic         wam_score
##  Test Wiki           :    67   Video Games  :55625   Min.   : 0.000
##  Naruto Wiki         :    59   Gaming       :36285   1st Qu.: 0.000
##  Harry Potter Wiki   :    54   Entertainment:29969   Median : 0.000
##  Pokemon Wiki        :    48   Creative     :25359   Mean   : 1.064
##  Minecraft Wiki      :    44   Fanon        :23055   3rd Qu.: 0.000
##  Final Frontier Wiki:    39   TV           :14075   Max.   :99.834
##  (Other)             :277484   (Other)      :93427
##  stats.nonarticles    users_1             users_5
##  Min.   :     -1   Min.   :     0.00   Min.   :    0.0
##  1st Qu.:     95   1st Qu.:     4.00   1st Qu.:    1.0
##  Median :    150   Median :     5.00   Median :    2.0
##  Mean   :    924   Mean   :    27.48   Mean   :   11.5
##  3rd Qu.:    256   3rd Qu.:     9.00   3rd Qu.:    4.0
##  Max.   :7725359   Max.   :127087.00   Max.   :32279.0
##
##     users_10           users_20            users_50
##  Min.   :    0.00   Min.   :    0.000   Min.   :    0.000
##  1st Qu.:    1.00   1st Qu.:    0.000   1st Qu.:    0.000
##  Median :    1.00   Median :    1.000   Median :    1.000
##  Mean   :    7.67   Mean   :    5.044   Mean   :    2.813
##  3rd Qu.:    3.00   3rd Qu.:    2.000   3rd Qu.:    1.000
##  Max.   :18682.00   Max.   :15113.000   Max.   :11067.000
##
##    users_100             bots                              birthDate
##  Min.   :   0.000   Min.   : 0.000   20:00, December 11, 2013 :   379
##  1st Qu.:   0.000   1st Qu.: 2.000   09:37, September 25, 2006:   348
##  Median :   0.000   Median : 4.000   21:24, February 22, 2018 :   224
##  Mean   :   1.756   Mean   : 4.135   15:07, October 4, 2016   :   168
##  3rd Qu.:   1.000   3rd Qu.: 5.000   21:27 22 feb 2018        :    73
##  Max.   :8521.000   Max.   :34.000   22:19 29 mar 2017        :    69
##                                      (Other)                  :276534
##            datetime.birthDate
##  2013-12-11 20:00:00:   380
##  2006-09-25 09:37:00:   348
##  2018-02-22 21:24:00:   224
##  2016-10-04 15:07:00:   168
##  2018-02-22 21:27:00:    73
##  2017-03-29 22:19:00:    69
```

```
##  (Other)            :276533
```

Los campos url y domain son identificadores de la wiki y deberían ser únicos. Aunque tenemos un repetido en el dominio cliff-side.wikia.com que trataremos más adelante.

Ahora podemos deducir que headline se refiere a una especie de subtítulo de la wiki. En cualquier caso, se trata de un campo de texto informativo para los usuarios de la wiki, pero que a nosotros no nos interesa.

Tenemos muchísimas wikis sin videos y también muchas sin imágenes.

Wikia page views dataset

Ahora procedemos a cargar los datos de visitas:

```
# Cargamos el juego de datos
wikis_pgv<-read.csv("data/20181113_wikia-page-views.csv",header=T,sep=",")
head(wikis_pgv)
```

```
##                                                  url visited_pages
## 1                          http://spellmagotm.wikia.com             0
## 2 http://2017-monster-energy-nascar-cup-series.wikia.com             0
## 3                   http://10low46japreligion.wikia.com             0
## 4                             http://de.bibel.wikia.com             0
## 5                        http://indigo-showdown.wikia.com             0
## 6                             http://animewiki2.wikia.com            41
##   total_views
## 1           0
## 2           0
## 3           0
## 4           0
## 5           0
## 6         560
```

Mostramos información descriptiva de estos datos:

```
str(wikis_pgv)
```

```
## 'data.frame':    278889 obs. of  3 variables:
## $ url       : Factor w/ 278888 levels "http://0002oifos.wikia.com",..: 225669 511 138 40364 122441 9030 5
##  $ visited_pages: int  0 0 0 0 0 41 0 0 5 0 ...
##  $ total_views  : int  0 0 0 0 0 560 0 0 47 0 ...
```

```
summary(wikis_pgv)
```

```
##                                        url         visited_pages
##  http://pl.6bp-6-batalion-pancerny.wikia.com:     2   Min.   :   0.00
##  http://0002oifos.wikia.com                 :     1   1st Qu.:   0.00
##  http://001-game-creator.wikia.com          :     1   Median :   0.00
##  http://001littlebighelp.wikia.com          :     1   Mean   :  12.36
##  http://007fanon.wikia.com                  :     1   3rd Qu.:   2.00
##  http://007goldeneye.wikia.com              :     1   Max.   :1000.00
##  (Other)                                    :278882
##   total_views
##  Min.   :       0
##  1st Qu.:       0
##  Median :       0
##  Mean   :    2892
```

```
##  3rd Qu.:      15
##  Max.   :20386167
##
```

Observamos que hay una wiki duplicada: `http://pl.6bp-6-batalion-pancerny.wikia.com/`. También observamos que la mayoría de la wikis (más de la mitad) no han tenido ni una sola visita a sus páginas en las últimas cuatro semanas. Es decir, podríamos considerar que estas wikis están muertas, puesto que ni siquiera usuarios de internet externos a la comunidad las visitan. En el otro extremo tenemos también otras wikis muy populares y así vemos como las medias que obtenemos de tanto páginas visitas como de visitas son mucho mayores que cero. Pero estas wikis que acumulan muchas visitas son escasas y no aparecen hasta más tarde del tercer cuartil de wikis.

Mostramos top 10 wikis con mayor número de visitas:

```
head(wikis_pgv[with(wikis_pgv, order(desc(wikis_pgv$total_views))), ], n = 10)
```

```
##                                      url visited_pages total_views
## 65555  http://oldschoolrunescape.wikia.com           999    20386167
## 156168            http://fallout.wikia.com           996    11916618
## 74539               http://dnd5e.wikia.com           676    11416101
## 28452            http://warframe.wikia.com           998    11165952
## 5484         http://elderscrolls.wikia.com           999    11160352
## 253171             http://naruto.wikia.com           998     9236970
## 246128 http://bokunoheroacademia.wikia.com           996     8660570
## 202431            http://reddead.wikia.com           978     7722284
## 228607           http://onepiece.wikia.com           993     7543366
## 273208             http://yugioh.wikia.com           996     7038509
```

Y top 10 wikis con mayor número de páginas visitadas:

```
head(wikis_pgv[with(wikis_pgv, order(desc(wikis_pgv$visited_pages))), ], n = 10)
```

```
##                                     url visited_pages total_views
## 24049         http://skylanders.wikia.com          1000      384642
## 24429    http://pokemon-uranium.wikia.com          1000      927726
## 32078  http://unsolvedmysteries.wikia.com          1000      265778
## 58515              http://fable.wikia.com          1000      283271
## 61960           http://terraria.wikia.com          1000      607245
## 73099         http://yokaiwatch.wikia.com          1000      858819
## 75246           http://gamelore.wikia.com          1000       11782
## 96655          http://pl.naruto.wikia.com          1000      336321
## 97769          http://saintsrow.wikia.com          1000      250143
## 104590       http://ru.grimdawn.wikia.com          1000      181918
```

Limpieza

Vamos a hacer limpieza de los datos que muestran valores raros o que no deberían estar: Empezamos por tratar que el número de imágenes sea negativo:

Fijaremos a 0 cuando stats.images sea inferior a 0.

```
wikis$stats.images[wikis$stats.images < 0] = 0
```

Fijaremos a 0 cuando stas.activeUsers sea inferior a 0. (Significa que no tenemos usuarios activos en esa wiki, pero no tiene sentido que tengamos valores menores que 0):

```
wikis$stats.activeUsers[wikis$stats.activeUsers < 0] = 0
```

Eliminamos wikis con stats.users o stats.nonarticles o stats.pages menores que cero, puesto que más bien representan que la wiki no tiene datos válidos (una wiki normal al menos debe tener un usuario registrado o una página):

```
invalid_wikis = wikis$stats.users < 0 | wikis$stats.pages < 0 | wikis$stats.nonarticles < 0
dim(wikis[invalid_wikis,]) # number of invalid wikis to delete
```

```
## [1]  3 32
```

```
wikis = wikis[-invalid_wikis,]
dim(wikis)
```

```
## [1] 277794     32
```

Después, vemos qué pasa con los duplicados por dominio:

```
wikis[duplicated(wikis$domain),]
```

```
##                               url       creation_date               domain
## 128184 http://cliffside.wikia.com 2018-08-18 06:39:50 cliff-side.wikia.com
##        founding_user_id headline hub      id lang language           name
## 128184          1824669           TV 1788785   en       en CliffSide Wiki
##        stats.activeUsers stats.admins stats.articles stats.discussions
## 128184                 1            1             56                 0
##        stats.edits stats.images stats.pages stats.users stats.videos
## 128184         859          173         393    15621017           10
##                 title         topic wam_score stats.nonarticles users_1
## 128184 CliffSide Wiki Entertainment         0               337       5
##        users_5 users_10 users_20 users_50 users_100 bots
## 128184       1        1        1        1         1    2
##                     birthDate  datetime.birthDate
## 128184 06:40, August 18, 2018 2018-08-18 06:40:00
```

```
#wikis[domain == "cliff-side.wikia.com", ] # Solo hay este dominio duplicado
# el dominio cliff-side.wikia.com está repetido. Eliminamos el último:
wikis = wikis[-duplicated(wikis$domain),]
dim(wikis)
```

```
## [1] 277793     32
```

Eliminamos también el duplicado que hemos visto para los datos de las visitas:

```
wikis_pgv = wikis_pgv[!duplicated(wikis_pgv$url),]
summary(wikis_pgv)
```

```
##                                 url         visited_pages
##  http://0002oifos.wikia.com       :     1   Min.   :   0.00
##  http://001-game-creator.wikia.com:     1   1st Qu.:   0.00
##  http://001littlebighelp.wikia.com:     1   Median :   0.00
##  http://007fanon.wikia.com        :     1   Mean   :  12.36
##  http://007goldeneye.wikia.com    :     1   3rd Qu.:   2.00
##  http://007-james-bond.wikia.com  :     1   Max.   :1000.00
##  (Other)                          :278882
##   total_views
##  Min.   :       0
##  1st Qu.:       0
##  Median :       0
```

```
##  Mean   :    2892
##  3rd Qu.:      15
##  Max.   :20386167
##
```

Transformación

En lugar de tener la fecha de creación del censo con formato fecha, que es un formato poco comparable y clasificable en intervalos, vamos a convertirlo a una nueva variable `age` que será la edad de la wiki en días:

```
wikis$datetime.birthDate = as.POSIXct(wikis$datetime.birthDate)
#str(wikis$datetime.birthDate)
wikis$age = as.integer(Sys.time() - wikis$datetime.birthDate)
summary(wikis$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      96     321    1176    1361    2185  736841
```

El valor máximo de 736808 corresponde al año 1 d.C., lo cual es imposible. Miramos cuántos valores de estos anómalos hay para age:

```
wikis[wikis$age > 365 * 22 , c('birthDate', 'datetime.birthDate', 'age')] # wikis con más de 22 años
```

```
##                   birthDate datetime.birthDate    age
## 201614 20:25, July 28, 0001   1-07-28 20:25:00 736841
```

Observamos que hay un error en la wiki con url: http://jrmime.wikia.com. La fecha en `birthDate` es incorrecta (año 1), mientras que la fecha de creación en realidad es 2013-03-11.

Lo arreglamos:

```
aux = wikis[wikis$age > 365 * 22 ,]
aux$age = as.integer(Sys.time() - as.POSIXct(aux$creation_date))
wikis[wikis$age > 365 * 22 ,] = aux
# Comprobamos de nuevo si hay algún otro caso raro:
wikis[wikis$age > 365 * 22 ,] # wikis con más de 22 años
```

```
##  [1] url                creation_date      domain
##  [4] founding_user_id   headline           hub
##  [7] id                 lang               language
## [10] name               stats.activeUsers  stats.admins
## [13] stats.articles     stats.discussions  stats.edits
## [16] stats.images       stats.pages        stats.users
## [19] stats.videos       title              topic
## [22] wam_score          stats.nonarticles  users_1
## [25] users_5            users_10           users_20
## [28] users_50           users_100          bots
## [31] birthDate          datetime.birthDate age
## <0 rows> (or 0-length row.names)
```

```
summary(wikis$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      96     321    1176    1358    2185    7507
```

Integración

Ahora vamos a unir los dos datasets de los cuales disponemos: Wikia Census y Wikia page views. Es lo podemos hacer añadiendo los datos de visitas al dataframe `wikis` que ya tenemos. Para ello debemos juntar ambos dataframes usando como columna identificadora común la columna `url`.

```
wikis_all = merge(wikis, wikis_pgv, by="url")
dim(wikis_all)
```

```
## [1] 266408     35
```

```
summary(wikis_all)
```

```
##                                 url                     creation_date
##  http://0002oifos.wikia.com       :     1   2006-03-16 13:42:45:   370
##  http://001-game-creator.wikia.com:     1   2004-11-11 23:33:14:    77
##  http://001littlebighelp.wikia.com:     1   2017-10-16 21:02:01:    12
##  http://007fanon.wikia.com        :     1   2006-10-28 11:16:22:    11
##  http://007goldeneye.wikia.com    :     1   2017-10-02 21:37:50:     8
##  http://007-james-bond.wikia.com  :     1   2018-02-08 13:53:15:     7
##  (Other)                          :266402   (Other)            :265923
##                          domain       founding_user_id
##  cliff-side.wikia.com      :     2   Min.   :       0
##  0002oifos.wikia.com       :     1   1st Qu.: 4893531
##  001-game-creator.wikia.com:     1   Median :25311196
##  001littlebighelp.wikia.com:     1   Mean   :19946785
##  007fanon.wikia.com        :     1   3rd Qu.:32232648
##  007goldeneye.wikia.com    :     1   Max.   :36902116
##  (Other)                   :266401   NA's   :35
##                                    headline             hub
##                                       :258992   Games    :100378
##  Alice Wiki                           :     2   Lifestyle: 62942
##  Assassin's Creed Wiki                :     2   TV       : 42839
##  Bakugan Wiki                         :     2   Books    : 21363
##  Bienvenidos a PokéMewtwo y PokéFantasy.:     2   Comics   : 13757
##  BioShock Wiki                        :     2   Movies   : 13436
##  (Other)                              :  7406   (Other)  : 11693
##        id               lang           language
##  Min.   :      1   en     :191596   en     :191596
##  1st Qu.: 643464   es     : 28618   es     : 28618
##  Median :1240829   ru     : 12298   ru     : 12298
##  Mean   :1121512   de     :  8741   de     :  8741
##  3rd Qu.:1668302   pl     :  7318   pl     :  7318
##  Max.   :1807291   fr     :  6558   fr     :  6558
##                    (Other): 11279   (Other): 11279
##                   name        stats.activeUsers   stats.admins
##  Test Wiki          :    60   Min.   :   0.000   Min.   :  0.000
##  Naruto Wiki        :    53   1st Qu.:   0.000   1st Qu.:  1.000
##  Harry Potter Wiki  :    46   Median :   0.000   Median :  1.000
##  Pokemon Wiki       :    46   Mean   :   0.813   Mean   :  1.483
##  Minecraft Wiki     :    42   3rd Qu.:   0.000   3rd Qu.:  1.000
##  Final Frontier Wiki:    39   Max.   :7311.000   Max.   :248.000
##  (Other)            :266122
##  stats.articles      stats.discussions   stats.edits
##  Min.   :      0.0   Min.   :    0.00   Min.   :       0
```

```
##  1st Qu.:      5.0   1st Qu.:    0.00   1st Qu.:     121
##  Median :     10.0   Median :    0.00   Median :     250
##  Mean   :    104.6   Mean   :    2.56   Mean   :    2482
##  3rd Qu.:     30.0   3rd Qu.:    1.00   3rd Qu.:     478
##  Max.   :1145898.0   Max.   :77051.00   Max.   :23302062
##                      NA's   :51310
##   stats.images       stats.pages       stats.users
##  Min.   :     0.0   Min.   :     -1   Min.   :      -1
##  1st Qu.:     2.0   1st Qu.:    103   1st Qu.:16129209
##  Median :    11.0   Median :    181   Median :16362464
##  Mean   :   190.1   Mean   :    949   Mean   :16221110
##  3rd Qu.:    41.0   3rd Qu.:    288   3rd Qu.:16543432
##  Max.   :784356.0   Max.   :7727191   Max.   :23922667
##
##   stats.videos                      title                   topic
##  Min.   :    0.00   Test Wiki          :    60   Video Games  :52417
##  1st Qu.:    0.00   Naruto Wiki        :    53   Gaming       :35770
##  Median :    0.00   Harry Potter Wiki  :    46   Entertainment:28205
##  Mean   :    6.76   Pokemon Wiki       :    46   Creative     :25021
##  3rd Qu.:    0.00   Minecraft Wiki     :    42   Fanon        :22258
##  Max.   :51799.00   Final Frontier Wiki:    39   TV           :13509
##                     (Other)            :266122   (Other)      :89228
##    wam_score       stats.nonarticles    users_1
##  Min.   : 0.0000   Min.   :     -1   Min.   :     0.00
##  1st Qu.: 0.0000   1st Qu.:     96   1st Qu.:     4.00
##  Median : 0.0000   Median :    156   Median :     5.00
##  Mean   : 0.9481   Mean   :    844   Mean   :    25.09
##  3rd Qu.: 0.0000   3rd Qu.:    257   3rd Qu.:     9.00
##  Max.   :99.8342   Max.   :7725359   Max.   :127087.00
##
##    users_5            users_10            users_20
##  Min.   :    0.00   Min.   :    0.000   Min.   :    0.0
##  1st Qu.:    1.00   1st Qu.:    1.000   1st Qu.:    0.0
##  Median :    2.00   Median :    1.000   Median :    1.0
##  Mean   :   10.44   Mean   :    6.975   Mean   :    4.6
##  3rd Qu.:    4.00   3rd Qu.:    3.000   3rd Qu.:    2.0
##  Max.   :32279.00   Max.   :18682.000   Max.   :15113.0
##
##    users_50           users_100             bots
##  Min.   :    0.000   Min.   :   0.000   Min.   : 0.000
##  1st Qu.:    0.000   1st Qu.:   0.000   1st Qu.: 2.000
##  Median :    1.000   Median :   0.000   Median : 4.000
##  Mean   :    2.573   Mean   :   1.606   Mean   : 4.167
##  3rd Qu.:    1.000   3rd Qu.:   1.000   3rd Qu.: 5.000
##  Max.   :11067.000   Max.   :8521.000   Max.   :34.000
##
##                      birthDate      datetime.birthDate
##  20:00, December 11, 2013 :   375   Min.   :1-07-28 20:25:00
##  09:37, September 25, 2006:   332   1st Qu.:2012-12-02 14:52:45
##  21:24, February 22, 2018 :   211   Median :2015-08-03 18:29:00
##  15:07, October 4, 2016   :   165   Mean   :2015-03-04 03:23:54
##  21:27 22 feb 2018        :    73   3rd Qu.:2017-12-27 03:25:15
##  22:19 29 mar 2017        :    67   Max.   :2018-09-17 19:34:00
##  (Other)                  :265185
```

```
##       age          visited_pages      total_views
##  Min.   :  96   Min.   :   0.00   Min.   :       0
##  1st Qu.: 361   1st Qu.:   0.00   1st Qu.:       0
##  Median :1237   Median :   0.00   Median :       0
##  Mean   :1387   Mean   :  12.47   Mean   :    2959
##  3rd Qu.:2212   3rd Qu.:   2.00   3rd Qu.:      15
##  Max.   :7507   Max.   :1000.00   Max.   :20386167
##
```

Resultados

Finalmente, vamos a producir un fichero con los datos ya limpiados, transformados e integrados:

```
write.csv(wikis_all, "output_data/20181019-wikia_census_and_page_views-clean.csv")
```

Referencias

Jimenez-Diaz, Guillermo, Abel Serrano, y Javier Arroyo. 2018. «A Wikia Census: Motives, Tools and Insights». En *Proceedings of the 14th International Symposium on Open Collaboration*, 2:1-2:6. OpenSym '18. New York, NY, USA: ACM. doi:10.1145/3233391.3233526.