22/23

Paquetes Avanzados con R

Master Big Data & Business Analytics

BRUNO URBAN ALFARO

UNED

Tabla de contenido

Preparación de database

Lo primero que hacemos es convertir

```
uknames <- as.data.table(ukbabynames)
usanames <- as.data.table(babynames)
```

Observamos los campos de cada base de datos y vemos que no son los mismos-

```
> head(uknames)
   year sex    name    n rank          nation
1: 1996   F  SOPHIE 7087    1 England & Wales
2: 1996   F   CHLOE 6824    2 England & Wales
3: 1996   F JESSICA 6711    3 England & Wales
4: 1996   F   EMILY 6415    4 England & Wales
5: 1996   F  LAUREN 6299    5 England & Wales
6: 1996   F  HANNAH 5916    6 England & Wales
> head(usanames)
   year sex      name    n       prop
1: 1880   F      Mary 7065 0.07238359
2: 1880   F      Anna 2604 0.02667896
3: 1880   F      Emma 2003 0.02052149
4: 1880   F Elizabeth 1939 0.01986579
5: 1880   F    Minnie 1746 0.01788843
6: 1880   F  Margaret 1578 0.01616720
```

Como queremos unir estas bases de datos, necesitamos que las variables coincidan, así que vamos a eliminar las que no coincidan (ya que al no tenerlas en la otra base de datos no podemos usar esa información para comparar uk y usa). Hemos añadido nation en los datos de USA también.

```
> head(uknames)
   year sex    name    n          nation
1: 1996   F  SOPHIE 7087 England & Wales
2: 1996   F   CHLOE 6824 England & Wales
3: 1996   F JESSICA 6711 England & Wales
4: 1996   F   EMILY 6415 England & Wales
5: 1996   F  LAUREN 6299 England & Wales
6: 1996   F  HANNAH 5916 England & Wales
> head(usanames)
   year sex      name    n nation
1: 1880   F      Mary 7065    USA
2: 1880   F      Anna 2604    USA
3: 1880   F      Emma 2003    USA
4: 1880   F Elizabeth 1939    USA
5: 1880   F    Minnie 1746    USA
6: 1880   F  Margaret 1578    USA
```

Ahora unimos las bases de datos para trabajar con todos los datos juntos:

```
allbabynames <- rbind(uknames,usanames)
```

Ahora que tenemos nuestra base de datos podemos comenzar el análisis.

Análisis

Nombres más comunes por nación

Lo primero es agrupar los datos por nación y nombre:

```
> popular_pais <-allbabynames %>%
+            group_by(nation, name) %>%
+            summarise(n = sum(n))
```

Lo ordenamos poniendo los nombres más populares en la cima:

```
popular_pais_orden <- popular_pais %>%
                                group_by(nation, name)%>%
                              arrange(desc(n)) %>%
                                mutate(rank = rank(-n, ties.method = "first"))
```

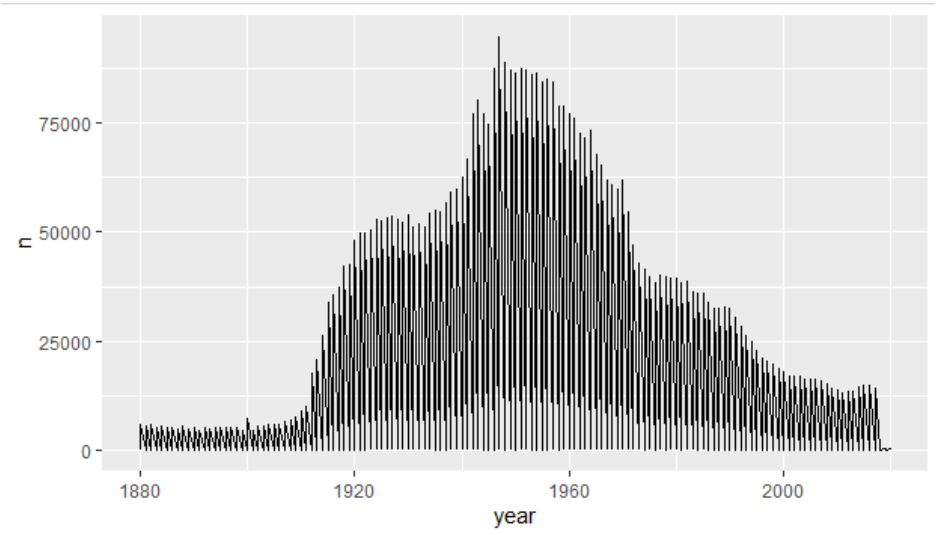
Por último, filtramos por naciones distintas para ver cuál es el más popular en cada una:

```
> popular_pais_orden %>%
+            distinct(nation)%>%
+ group_by(name)
# A tidytable: 4 x 2
# Groups:      name
  nation           name
  <chr>            <chr>
1 USA              James
2 England & Wales  JACK
3 Scotland         David
4 Northern Ireland Jack
```

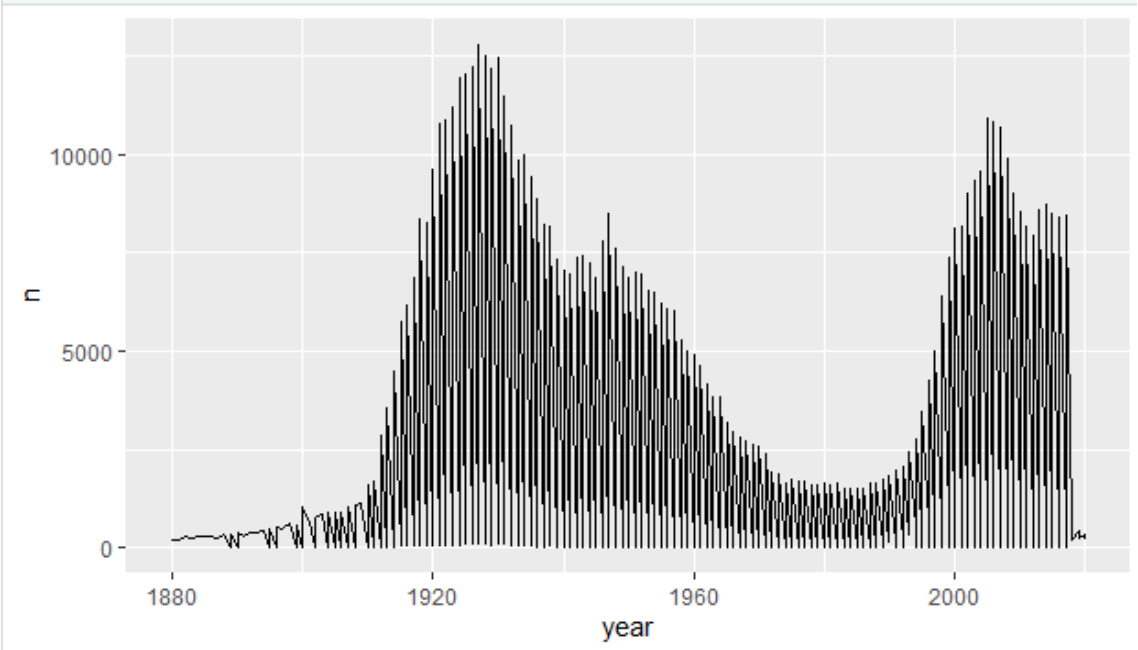
Gracias a este anális concluimos que Jack es el nombre más popular tanto en Inglaterra y Gales como en Irlanda.

Para visualizar esta popularidad, vamos a crear un gráfico donde analizaremos si esta cantidad ha sido siempre igual en la historia.

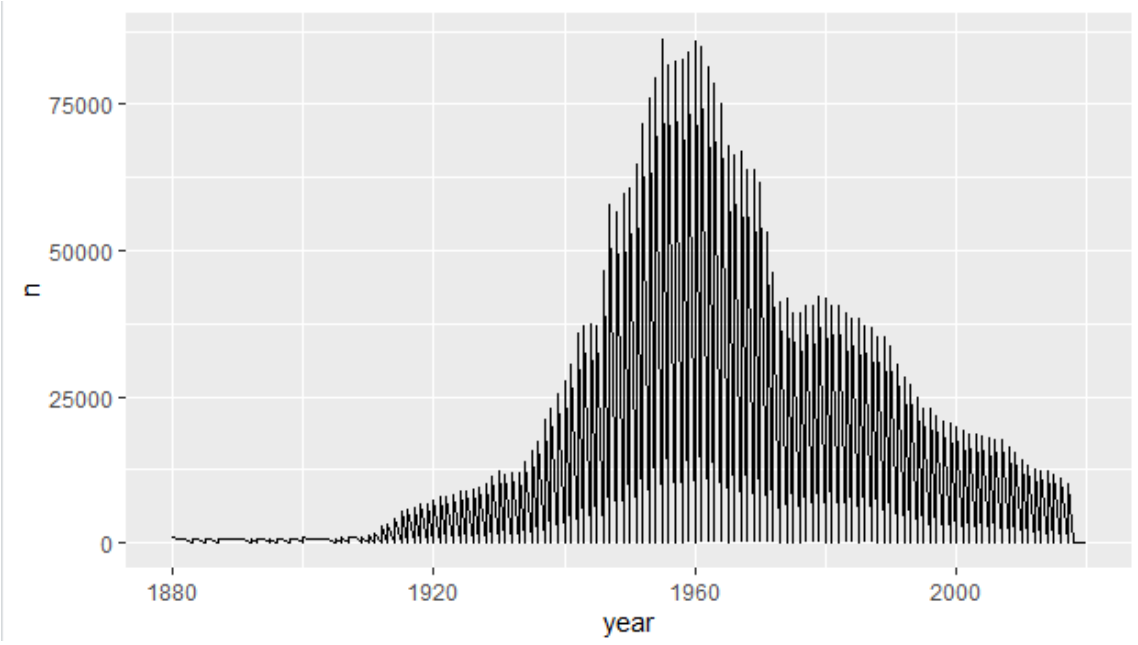
Primero vamos a ver el nombre “James”:



Claramente la popularidad de éste tuvo su pico entre los años 40 y 60, disminuyendo drásticamente hasta nuestra época.



Curiosamente, “Jack” tuvo un pico tirando más a los años 30, y otro entrando en los 2000.



“David” muestra un comportamiento más parecido a “James”.

Viendo estos gráficos se observa que la popularidad de “James” viene de haber sido un nombre común durante un tiempo más prolongado, ya que si nos fijamos en picos, “David” y “Jack han sido nombres con más portadores a la vez.

Nombres más comunes por sexo

Siguiendo los pasos previos

```
popular_sex <-allbabynames %>%
                    group_by(sex, name) %>%
                  summarize(n = sum(n))
```

```
>  popular_sex_orden <- popular_sex %>%
+                                        group_by(sex, name)%>%
+                                      arrange(desc(n)) %>%
+                                        mutate(rank = rank(-n, ties.method = "first"))
```

Podemos ver que el nombre más popular de chico es James y de chica Mary.

Pero vamos más allá y vemos los nombres más populares en general:

	sex	name	n
1	M	James	5191487
2	M	John	5146528
3	M	Robert	4834453
4	M	Michael	4377475
5	F	Mary	4126357
6	M	William	4120845
7	M	David	3653965
8	M	Joseph	2612637
9	M	Richard	2571751
10	M	Charles	2390094

Al sacar los primeros campos de los nombres más populares vemos que de los 10 primeros, únicamente uno es de chica. Esto puede significar que hay más variedad en los nombres de chica. Vamos a comprobarlo.

```
> nombres_chico <-allbabynames %>%
+       filter(sex=="M")%>%
+  distinct(name)
> nombres_chica <-allbabynames %>%
+       filter(sex=="F")%>%
+       distinct(name)
> count(nombres_chico)
# A tidytable: 1 × 1
      n
  <int>
1 74131
> count(nombres_chica)
# A tidytable: 1 × 1
      n
   <int>
1 114380
```

Efectivamente, hay mucha más variedad de nombres de chica (114.380 contra 74.131) por lo que tiene sentido que los más populares sean predominantemente masculinos, ya que tienen menos opciones.

Nombres más comunes a lo largo del tiempo

Para analizar los nombres más populares a lo largo de la historia vamos a realizar unos pasos similares a los previos

```
> nombres_tiempo <- allbabynames %>%
+       group_by(nation, year, name) %>%
+       summarize(n = sum(n))
>
> nombres_tiempo_orden <- nombres_tiempo %>%
+      group_by(year, name) %>%
+      arrange(desc(n))%>%
+      mutate(rank = rank(-n, ties.method = "first"))
> nombres_tiempo_orden <- nombres_tiempo %>%
+ group_by(year, name)%>%
+ arrange(desc(n))%>%
+ mutate(rank = rank(-n, ties.method = "first"))
> nombres_tiempo_orden %>%
+ distinct(year)%>%
+ group_by(name)

# A tidytable: 141 × 2
# Groups:      name
    year name
   <dbl> <chr>
 1  1947 Linda
 2  1948 Linda
 3  1957 Michael
 4  1949 Linda
 5  1956 Michael
 6  1958 Michael
 7  1954 Michael
 8  1955 Michael
 9  1946 James
10  1951 James
```

Nombres más comunes unisex

Previamente creamos unos filtros para ver los nombres de chico y chica, así que ahora vamos a hacer una intersección con los que coinciden para analizar los nombres neutrales.

```
nombre_unisex <- merge(nombre_chico, nombre_chica, by=c("name", "year"), all=FALSE)
```

Vemos que se ponen casi un 10% de nombres unisex

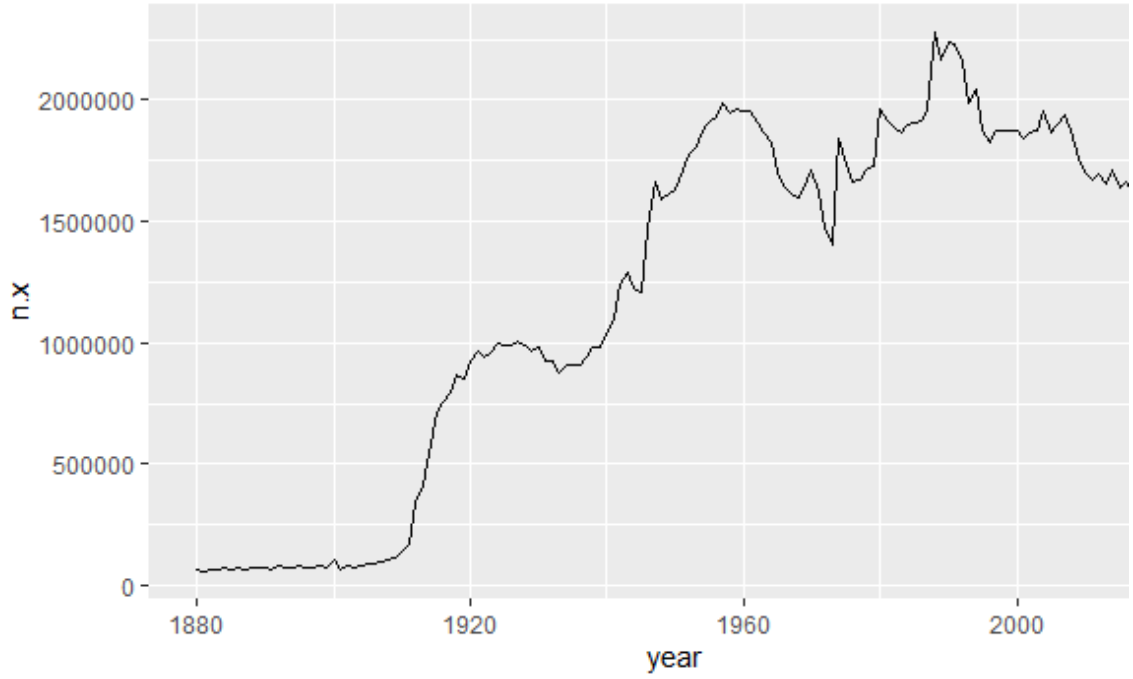
```
> ratio_neutral <- count(nombre_unisex)/count(allbabynames)*100
> ratio_neutral
          n
1: 9.788627
```

Podemos preguntarnos si los nombres unisex son más populares en este punto de la historia en el que hay un movimiento mayor de igualdad de género y de romper con los roles. Para ello vamos a comparar la cantidad desde 1947 hasta ahora.

Usamos la función aggregate para sumar los nombres unisex por años

```
unisex_agregados <- aggregate(n.x ~ year, data = nombre_unisex, sum)
```

```
> unisex_agregados%>%
+ ggplot(aes(x = year, y = n.x)) +
+  geom_line()
```



Si parece que han aumentado la cantidad de nombres unisex desde el siglo XX, pero a partir de los 60 no es algo realmente significativo.