Análisis y clasificación de nombres por género

Parte 2. Informe

Introducción

El objetivo de este informe fue realizar un análisis del listado de nombres registrados en el Registro Nacional de las Personas entre los años 1922 y 2015, agrupados por nombre y año, a fin de tener una primera aproximación y entendimiento de estos datos.

Como segundo objetivo se intentó determinar qué asociaciones y/o patrones de interés se podrían desprender a partir de este análisis.

Métodos y herramientas

La estructura del dataset es la siguiente:

Variable	Tipo	Descripción
nombre	caracter	Nombre registrado
cantidad	numérico	Cantidad del nombre registrado para un año
anio	numérico	Año del registro (entre 1922 y 2015)

El dataset se procesó en 3 etapas:

- chequeo y limpieza de datos. El procesamiento inicial se hizo en R, utilizando librerías para manipulación de dataset (dplyr, reshape2)

- manejo de datos y agregación. Para esto decidí usar MySQL (Workbench CE for Mac OS X version 5.2.47) por el manejo de datos y por las funciones de agrupación que brinda SQL
  Los datos fueron cargados mediante un loader y luego manipulados con scripts sql

- gráficos sobre conteo de nombres y caracteres. Este procesamiento se hizo en R utilizando la librería ggplot2

Resultados

Como un primer paso previo al análisis se analizó el dataset por posibles anomalias o errores de tipeo.
Se analizó por posibles datos faltantes, comprobando para los tres campos ausencia de NA:

	variable	perc_na
1	nombre	0
2	cantidad	0
3	anio	0

Analizando el campo nombre, se encontraron distintos tipos de errores y determinados ruidos que podían llegar a alterar el posterior análisis.

- anotaciones entre paréntesis, caracteres especiales y números: 960 registros. Estas anotaciones y números fueron removidos del dataset y aquellos casos donde el campo quedó nulo se removieron. Algunos siguientes ejemplos:

	nombre
1	Mercedes Nati9vidad
2	Aida Segunda Y0landa
3	Encarnacion (presunto 2843868)
4	Carlos 2°
5	Cecilio(presunto 7494687)
6	Juan Bautista (presunto 5639392)
7	Dominga(presunto 94121378)
8	Angelica(presunto3083950
9	Jorge 0legario
10	09/06/2010
11	Olga (Presunta 13380806)
12	Mario(presunto31678439)
13	Andres Horacio0.
14	Salvador(presunto12357774)
15	Eusebio (h)
16	Victoria (presunto)
17	Martina (presunta 4144699)
18	Guillermo (presunto 6602488)
19	Filomena(presunto 4747068)
20	Maxima(presunto10375842)

- Caracteres únicos y abreviaturas. Tanto los caracteres únicos solos como caracteres únicos seguidos de punto se consideraron como abreviaturas y se eliminaron de los registros. Se encontraron un total de 3433 registros como los siguientes:

	nombre
1	Claudia T
2	Emilia B
3	Epifanio de J
4	Amando G
5	Josefa C
6	Lidia B
7	Lucas E
8	Maria Nelly D R
9	Pascuala Del C
10	Ines E

Los datos se convirtieron a minúsculas y se bajaron a un archivo .csv para ser manipulados de manera más fácil por medio de MySQL.
Así se obtuvieron algunos datos de interés:

- Años donde se registraron mayor cantidad de registro de nombres:

registros	anio
1351210	1993
1327782	1994
1257479	1992
1245813	1991
1228279	1995

- Años donde se registraron menor cantidad de registro de nombres (desde 1978):

registros	anio
450370	2011
546042	2010
583182	2009
696160	2008
707364	2007
747118	2002
750246	2006
784934	2005
791107	2001

- Nombres con mayor cantidad de registros por década:

Década 1920	
Nombre	Registros
maria	4310
jose	3010
rosa	2961
maria esther	2145
antonio	2085

Década 1930	
Nombre	Registros
juan carlos	14245
maria	8857
jose	8158
rosa	7255
maria esther	7007

Década 1940	
Nombre	Registros
juan carlos	52389
ana maria	38741
carlos alberto	31701
miguel angel	29483
maria cristina	26507

Década 1950	
Nombre	Registros
juan carlos	69486
miguel angel	56889
ana maria	51998
maria cristina	49544
carlos alberto	44966

Década 1960	
Nombre	Registros
juan carlos	57505
miguel angel	51170
jose luis	42810
carlos alberto	38561
norma beatriz	30697

Década 1970	
Nombre	Registros
miguel angel	44200
juan carlos	41470
jose luis	34915
carlos alberto	34437
maria laura	29183

Década 1980	
Nombre	Registros
miguel angel	37243
maria laura	35927
juan manuel	35708
jose luis	33100
juan pablo	33065

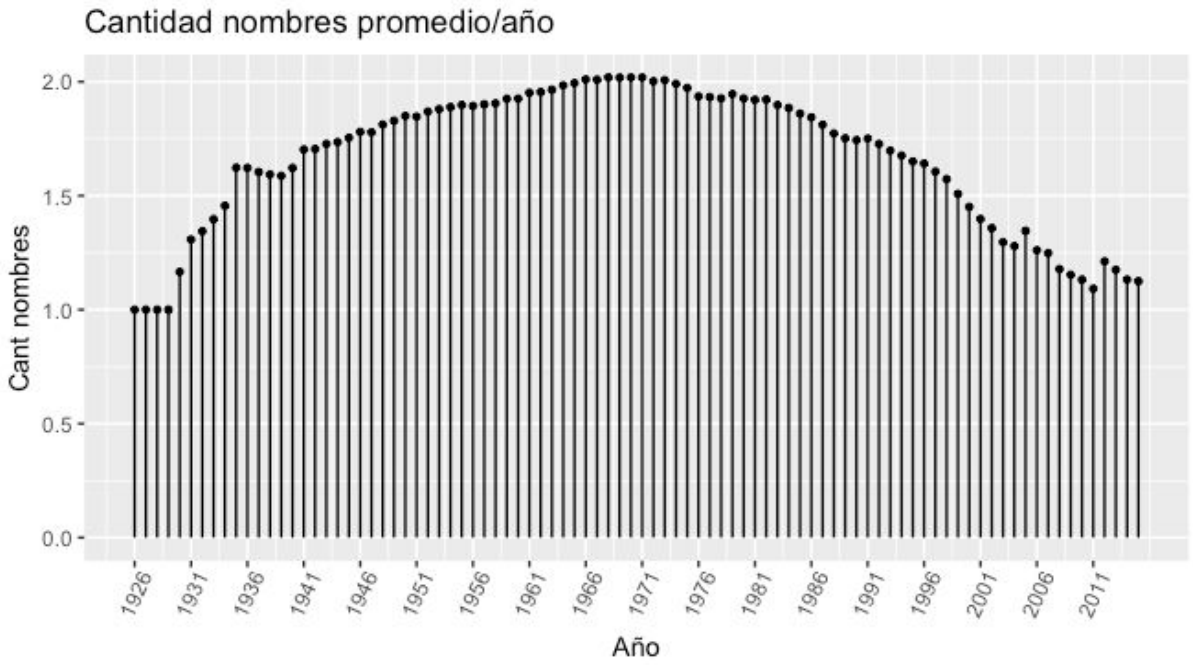
Década 1990	
Nombre	Registros
maría belén	46299
maria florencia	31292
juan manuel	30158
juan ignacio	25815
camila	25246

Década 2000	
Nombre	Registros
valentina	35357
martina	28950
santiago	25122
juan ignacio	23596
joaquin	22959

Década 2010	
Nombre	Registros
benjamin	23186
martina	17109
isabella	16539
bautista	15389
catalina	15382

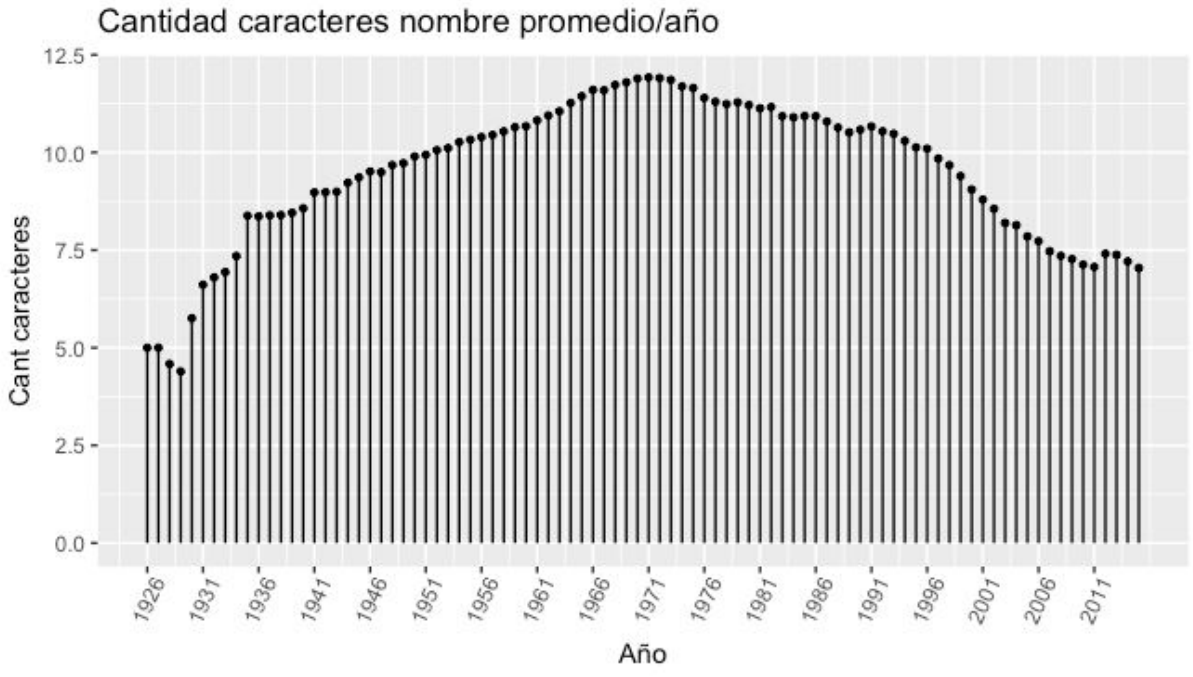
Analizando las tablas, se pueden ver nombres como Juan Carlos, Miguel Angel y Ana Maria que por varias décadas estuvieron entre los 5 nombres más registrados, contrario a las últimas décadas donde los nombres no se mantienen en este ranking durante más de una década.

- Cantidad de nombres promedio por año. Se tomaron para cada año los nombres con una cantidad de registros significativa (>500) y se calculó la cantidad de nombres promedio para analizar si a lo largo de los años se elegían nombres únicos, segundos nombres, etc:



Como puede verse en las décadas de los ‘60s y ‘70s el promedio de nombres se acerca mas a 2 por registro, y disminuye hacia los años extremos hasta quedar cercano a 1.

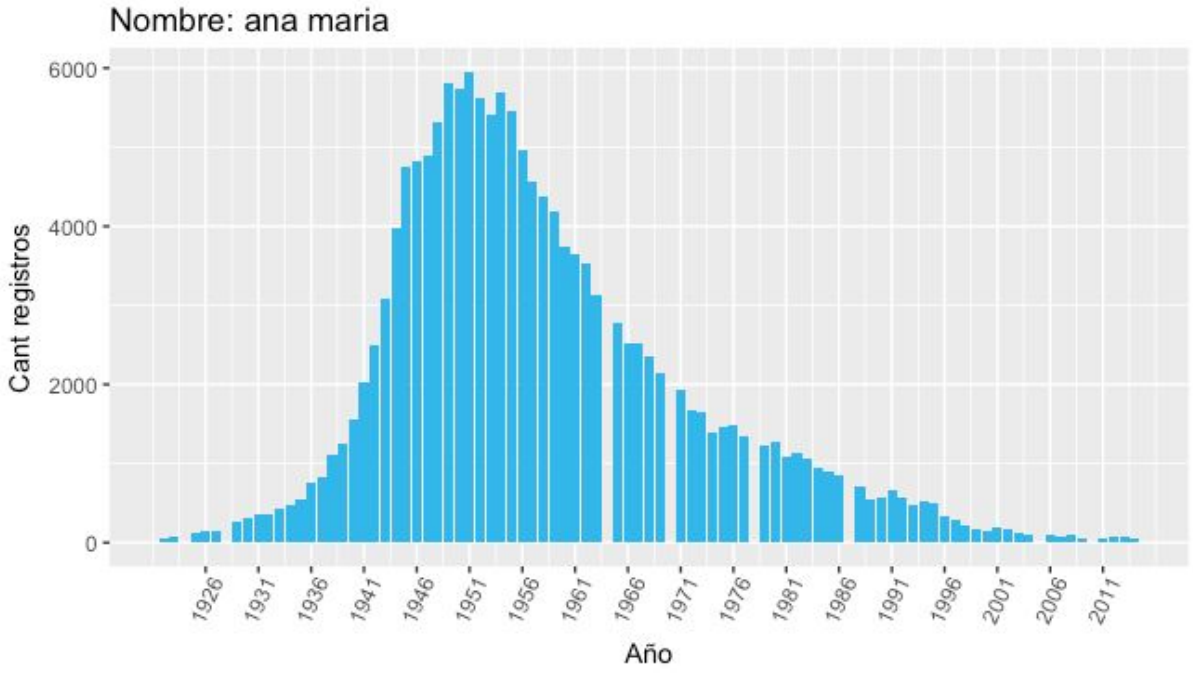
- Cantidad de caracteres por nombre promedio por año. Se tomaron para cada año los nombres con una cantidad de registros significativa (>500) y se calculó la cantidad de caracteres promedio del nombre completo para analizar si a lo largo de los años se elegían nombres completos de mayor longitud:



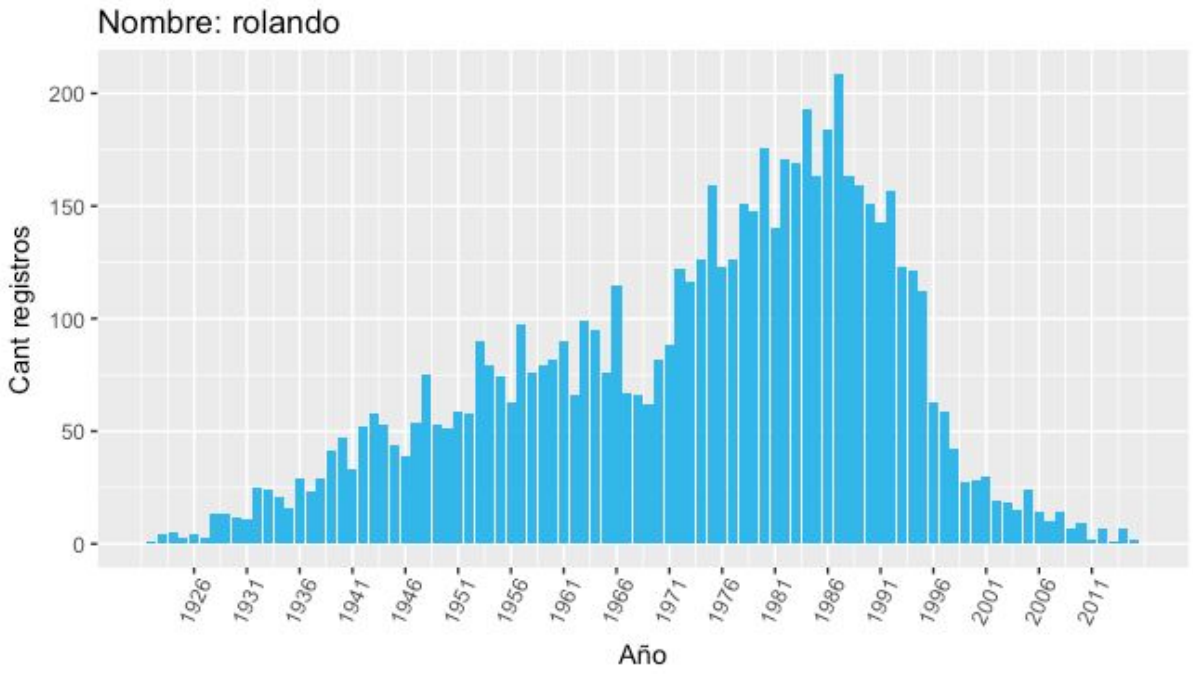
Al igual que con el promedio de nombres, en las décadas de los ‘60s y ‘70s la cantidad de caracteres promedio aumenta.
Nota: un paso más fino en este análisis hubiera sido calcular el promedio de caracteres por nombre por persona

- Asociación con personas / personajes argentinos

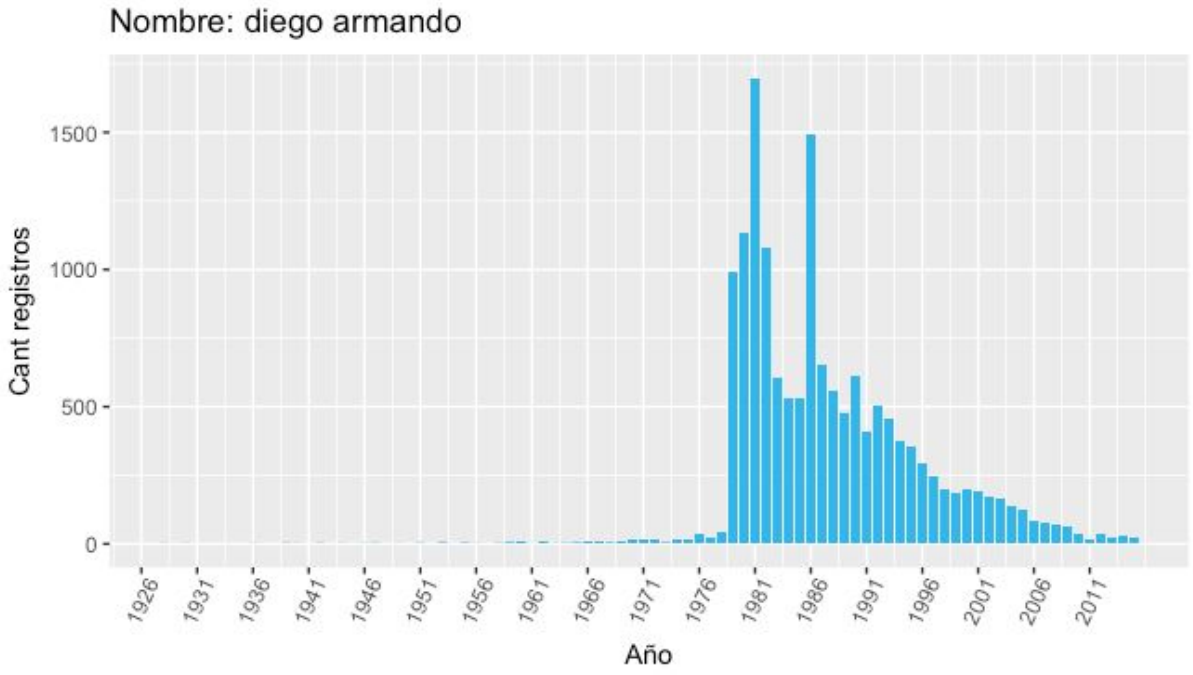
  Ana Maria: década del ‘50, el aumento de registros con este nombre puede relacionarse con telenovelas como “Teleteatro del suspenso” y “Como te quiero, Ana” que tuvieron como protagonista a la actriz Ana Maria Campoy



  Rolando: década del ‘70. Se observa a partir de principios de la década coincidente con la emisión de la telenovela “Rolando Rivas, taxista” el crecimiento en la cantidad de registros del nombre



Diego Armando. Coincidente con su inicio y participación en Argentinos Juniors entre 1976 y 1981, se puede observar el aumento de registros con este nombre, asi como tambien el pico que se ve en el año 1986, año que ganó el mundial de México con el equipo argentino



Referencias

Historia de la telenovela argentina:
https://es.wikipedia.org/wiki/Historia_de_la_telenovela_argentina