Análisis al dataframe arbuthnot

Alejandro Zavala

2023-08-09

Contents


```
# Clear environment
rm(list = ls())
# Call libraries
library("knitr")
library("ggplot2")
library("statsr")
```

{FALSE} ## Loading required package: BayesFactor

{FALSE} ## Loading required package: coda

{FALSE} ## Loading required package: Matrix

{FALSE} ## ************ ## Welcome to BayesFactor 0.9.12-4.4. If you have questions, please contact Richard Morey (richarddmorey@gmail.com). ## ## Type BFManual() to open the manual. ## ************

```
library("tidyverse")
```

{FALSE} ## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 -- ## v dplyr 1.1.2 v readr 2.1.4 ## v forcats 1.0.0 v stringr 1.5.0 ## v lubridate 1.9.2 v tibble 3.2.1 ## v purrr 1.0.1 v tidyr 1.3.0

{FALSE} ## -- Conflicts -- tidyverse_conflicts() -- ## x tidyr::expand() masks Matrix::expand() ## x dplyr::filter() masks stats::filter() ## x dplyr::lag() masks stats::lag() ## x tidyr::pack() masks Matrix::pack() ## x tidyr::unpack() masks Matrix::unpack() ## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

Descripción del análisis

Se hara un pequeño analisis a la tabla "arbuthnot" proveniente de la libreria **statsr**. Este dataframe contiene la cantidad de niños y niñas nacidos en Londres de 1629 a 1710

```
data(arbuthnot) # Load dataframa
kable(head(arbuthnot)) # Show few records
```

year	boys	girls
1629	5218	4683
1630	4858	4457
1631	4422	4102
1632	4994	4590
1633	5158	4839
1634	5035	4820

Explorando el dataframe

Describir dimensiones del dataframe

```
dim(arbuthnot) # Dimension of an object in R
```

```
## [1] 82  3
```

```
dim(arbuthnot)[1] # Total records of dataframe
```

```
## [1] 82
```

```
dim(arbuthnot)[2] # Total fields of dataframe
```

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

Donde el primer elemento denota la cantidad de registros totales y el segundo denota la cantidad de campos que contiene

Otra función interesante, es names que permite ver las variables del dataframe

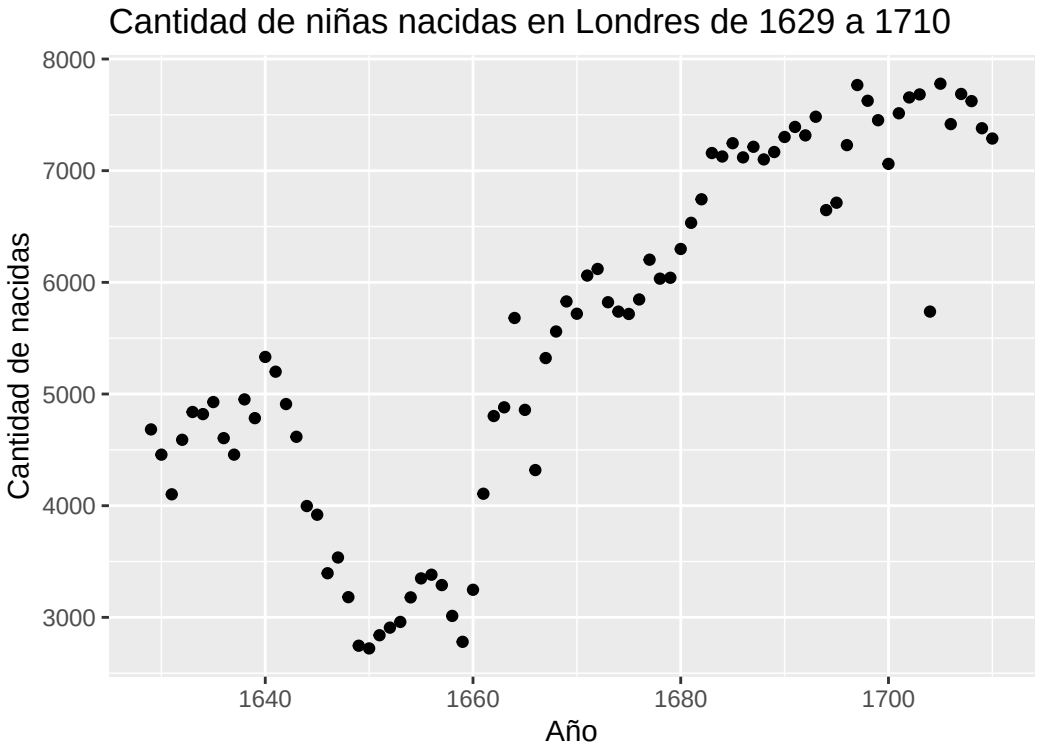
```
names(arbuthnot)
```

```
## [1] "year"  "boys"  "girls"
```

Visualización de los datos

Veamos ahora un diagrama de dispersion de la cantidad de niñas nacidas en Londres de 1629 a 1710

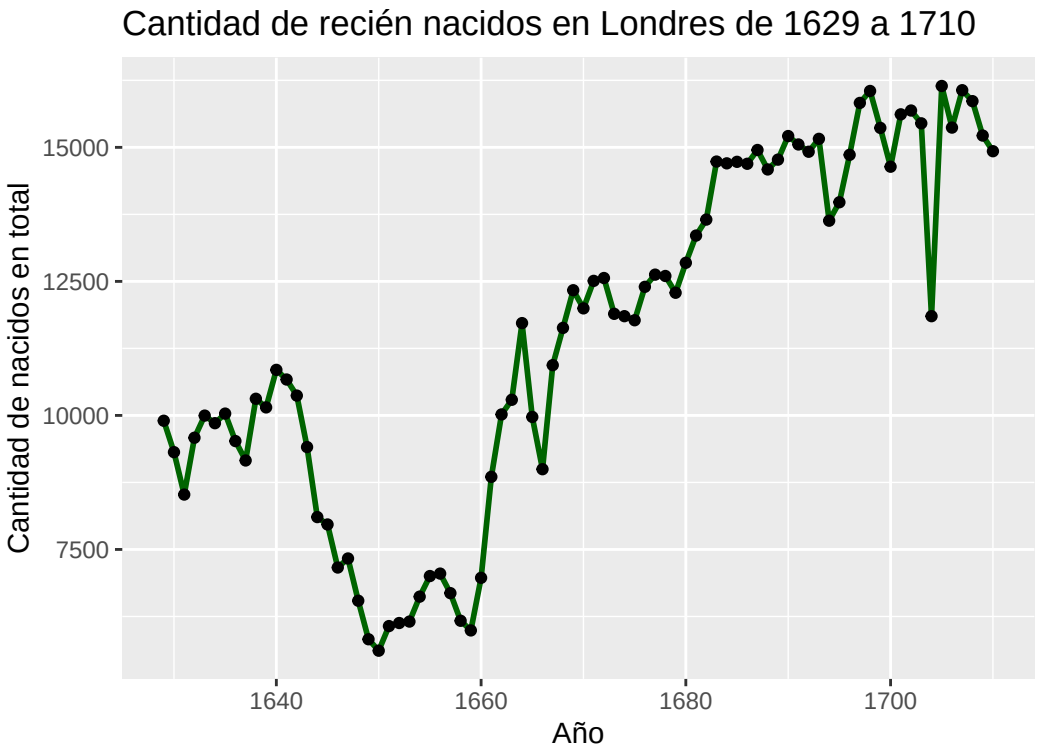
```
ggplot(data = arbuthnot, aes(x = year, y = girls)) +
  geom_point() +
  xlab("Año") +
  ylab("Cantidad de nacidas") +
  ggtitle("Cantidad de niñas nacidas en Londres de 1629 a 1710")
```



Construyamos un gráfico donde se muestre la cantidad total de niños y niñas nacidas en Londres de 1629 a 1710

```
arbuthnot$total <- arbuthnot$boys + arbuthnot$girls

ggplot(data = arbuthnot, aes(x = year, y = total)) +
  geom_line(size=1,color="darkgreen") +
  geom_point() +
  xlab("Año") +
  ylab("Cantidad de nacidos en total") +
  ggtitle("Cantidad de recién nacidos en Londres de 1629 a 1710")
```



Agregando nuevas variables

Veamos si hay mas niños o niñas por años, creando una variable que sea verdadera cuando haya mas niños que niña en cada año

```
arbuthnot$more_boys <- arbuthnot$boys > arbuthnot$girls
kable(head(arbuthnot))
```

year	boys	girls	total	more_boys
1629	5218	4683	9901	TRUE
1630	4858	4457	9315	TRUE
1631	4422	4102	8524	TRUE
1632	4994	4590	9584	TRUE
1633	5158	4839	9997	TRUE
1634	5035	4820	9855	TRUE

```
agg_boys_girls <- arbuthnot %>% group_by(more_boys) %>% summarise(total_count=n())
kable(agg_boys_girls) # Similar in SQL select more_boys from arbuthnot group by more_boys
```

more_boys	total_count
TRUE	82