

TRABALLAR CON DATOS DE XEITO LIMPO E EFICIENTE CO UNIVERSO *tidyverse*

María José Ginzo Villamayor

Universidade de Santiago de Compostela
Centro de Investigación e Tecnoloxía Matemática de Galicia

16 de outubro de 2025



DEPARTAMENTO DE ESTATÍSTICA,
ANÁLISE MATEMÁTICA E OPTIMIZACIÓN



CENTRO DE INVESTIGACIÓN
E TECNOLOXÍA MATEMÁTICA
DE GALICIA



Contidos

- » Introducción
- » O ciclo da análise de datos con R
- » Fluxo tidyverse
- » Aplicación práctica
- » Referencias bibliográficas

Obxectivos da análise

- Mostrar o potencial de **R** e o ecosistema `tidyverse` para traballar con datos xeográficos e sociodemográficos.
- Aplicar técnicas de limpeza e visualización a datos reais do **IGE**.
- Explorar a distribución e evolución temporal de apelidos e nomes en Galicia.

O Tidyverse

- O `tidyverse` é un conxunto integrado de paquetes de **R** deseñados para a **ciencia de datos**.
- Todos comparten unha mesma filosofía, gramática e estruturas de datos, o que facilita un fluxo de traballo coherente e reproducible.
- Paquetes principais:
 - `dplyr` – manipulación de datos
 - `ggplot2` – visualización
 - `tidyr` – reestruturación de datos
 - `readr` – importación de ficheiros
 - `stringr`, `forcats`, `lubridate`, `purrr`, `tibble`
- O seu obxectivo é facer a análise de datos **máis sinxela, consistente e divertida**.

0 Tidyverse



R packages for data science

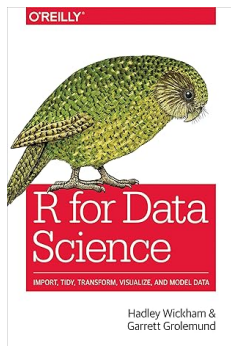
The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

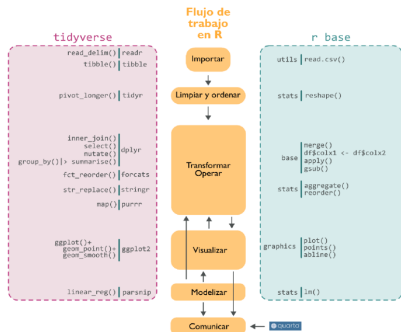
- <https://tidyverse.org/>

Aprender Tidyverse



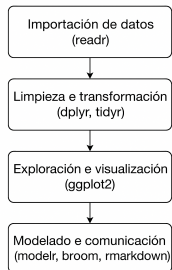
Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
Disponible en: <https://r4ds.hadley.nz>

Flujo de trabajo en R



- **Importar datos:** readr, tibble
- **Limpar e ordenar:** tidyr, dplyr
- **Transformar e operar:** purrr, forcats
- **Visualizar:** ggplot2
- **Modelizar:** parsnip
- **Comunicar:** Quarto

Fluxo resumido: tidyverse



Reflexión

Fluxo fundamental en análise de datos con R, combinando limpeza, exploración e modelado.

Distribución de apellidos en Galicia

Un exemplo real de análise con R

Os apellidos poden empregarse como **fonte de información** para caracterizar poboacións xa que a análise do **patrón dun apelido** proporciona información sobre tendencias de movementos de poboación ao longo do tempo.

Distribución de apellidos en Galicia

Un exemplo real de análise con R



Reión de apellidos en Galicia.



Sousa Fernández, X.; Crujeiras, R.M. and Ginzo-Villamayor, M.J. (2025). *Surname Distribution in Galicia: Population Structure, Dialects, and History. Names: A Journal of Onomastics*, 73(2), 59-75.

- Base de datos: padrón municipal de habitantes.
- Preprocesado con `tidyverse` `dplyr` e `stringr`.
- Análise exploratoria e visualización con `ggplot2`.
- Identificación de apellidos máis frecuentes por concello.
- Representación xeográfica e nube de apellidos...

Distribución de apellidos en Galicia

Descrición das variables do conxunto de datos

Variable	Significado	Tipo / Exemplo
cpro	Código da provincia de residencia	Numérico – ex. 15 (A Coruña)
cmun	Código do concello de residencia	Numérico – ex. 15030
cpron	Código da provincia de nacemento	Numérico – ex. 32 (Ourense)
cmunn	Código do concello de nacemento	Numérico – ex. 32054
part1	Partícula do apelido (p. ex. “del”, “de”)	Texto – ex. <i>del</i> en “del Río”
ape1	Primeiro apelido	Texto – ex. <i>Muñiz</i>
sexo	Sexo da persoa (1 = home, 6 = muller)	Numérico – 1 ou 6
idade	Idade no ano 2000	Numérico – ex. 45
cuenta	Número de persoas con esas características	Numérico – ex. 12

Preparación de datos con *tidyverse*

Código R

```
datos <- datos %>%  
  mutate(  
    part1 = replace_na(part1, ""),  
    ape1 = replace_na(ape1, ""),  
    apelido = paste0(part1, if_else(part1 == "", "", " "),  
    ape1),  
    apelido = str_squish(apelido)  
  )
```

- Xuntamos partícula e apelido nunha columna.
- Eliminamos valores NA e espazos extras.
- Boa práctica: `mutate()` para novas columnas limpas.
- Preparamos datos para agregación e visualización.

Agrupación por concello

Código R

```
datos <- datos %>%  
  mutate(  
    cpro = str_pad(as.character(cpro), 2, pad = "0"),  
    cmun = str_pad(as.character(cmun), 3, pad = "0"),  
    codigoine = str_c(cpro, cmun)  
  ) %>%  
  group_by(codigoine, apelido) %>%  
  summarise(cuenta = sum(cuenta, na.rm = TRUE)) %>%  
  ungroup()
```

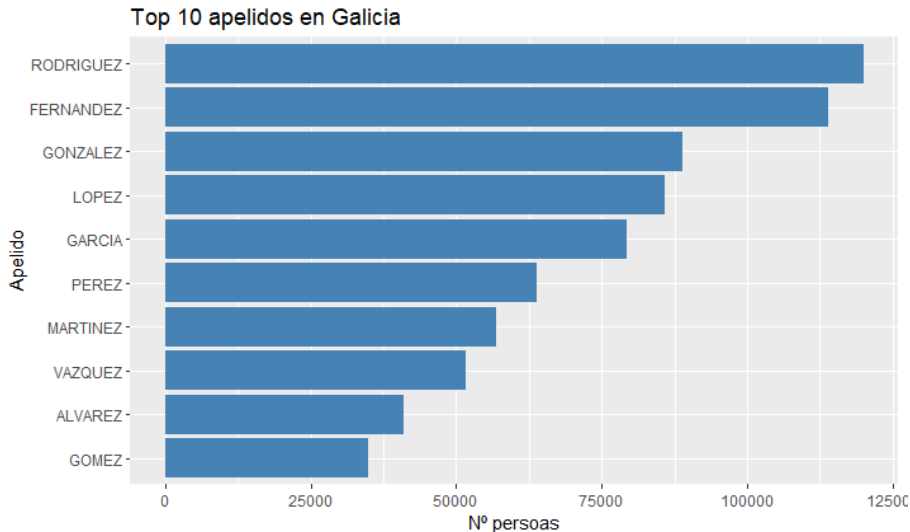
- Creamos códigos únicos de 5 dígitos para concellos.
- Agrupamos por concello e apelido para sumar persoas.
- Preparación esencial antes de unir cun mapa.

Top 10 apellidos en Galicia

- Código R:

```
datos %>%  
  group_by(apellido) %>%  
  summarise(total = sum(cuenta, na.rm = TRUE)) %>%  
  arrange(desc(total)) %>%  
  slice_head(n = 10) %>%  
  ggplot(aes(x = reorder(apellido, total), y = total)) +  
  geom_col(fill = "steelblue") +  
  coord_flip() +  
  labs(title = "Top 10 apellidos en Galicia",  
       x = "Apellido", y = "Nº persoas")
```

Top 10 apellidos en Galicia



Distribución do apelido **Rodríguez** en Galicia (ano 2000)

- `datos_galicia` obxecto da clase `sf`.

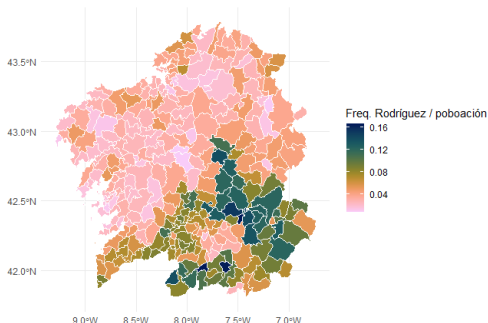
```
datos_galicia <- esp_muni_galicia %>%  
  left_join(datos, by = c("LAU_CODE" = "codigoine"))
```

- Código R:

```
rod_galicia_sf <- datos_galicia %>%  
  filter(apellido == "RODRIGUEZ") %>%  
  group_by(codigoine, pob2000) %>%  
  summarise(total = sum(cuenta, na.rm = TRUE)) %>%  
  ungroup() %>%  
  mutate(freq = total / pob2000) # frecuencia relativa
```

- Ten unha xeometría. Podemos pintar.

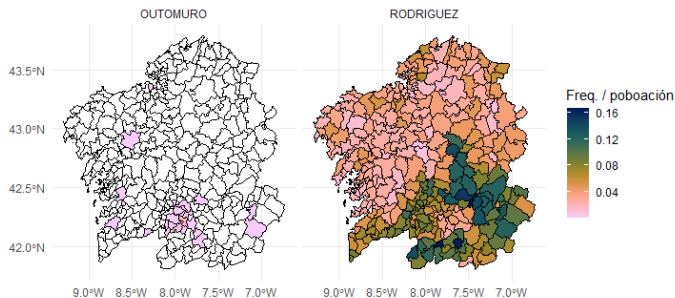
Distribución do apelido **Rodríguez** en Galicia (ano 2000)



- Cada concello coloreado segundo número de persoas.
- Escala de cores scico para mellor percepción.
- Frecuencia relativa: $\text{total} / \text{pob2000}$.
- Boa práctica: `geom_sf()` para mapas de polígonos.

Distribución de apellidos en Galicia

Comparación xeográfica de apellidos



Interpretación

RODRÍGUEZ presenta unha ampla distribución por toda Galicia, típica de apellidos moi comúns, mentres que **OUTOMURO** se concentra principalmente en zonas rurais de Ourense, reflectindo unha orixe máis localizada e patrimonial.

Distribución de nomes en Galicia

Un exemplo real de análise con R

Os nomes de pila tamén poden empregarse como **fonte de información** para caracterizar poboacións, xa que a análise do **patrón de nomeamento** proporciona datos sobre preferencias culturais, influencias históricas e cambios nas tendencias de elección de nomes ao longo do tempo.

Distribución de nomes en Galicia

Descrición das variables do conxunto de datos

Variable	Significado	Tipo / Exemplo
cpro	Código da provincia de residencia	Numérico – ex. 15 (A Coruña)
cmun	Código do concello de residencia	Numérico – ex. 15030
cpron	Código da provincia de nacemento	Numérico – ex. 32 (Ourense)
cmunn	Código do concello de nacemento	Numérico – ex. 32054
nomb	Nome	Texto – ex. <i>María</i>
sexo	Sexo da persoa (1 = home, 6 = muller)	Numérico – 1 ou 6
idade	Idade no ano 2000	Numérico – ex. 45
cuenta	Número de persoas con esas características	Numérico – ex. 12

Resumo de nomes en Galicia

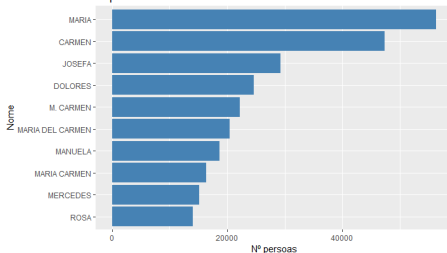
- Número de nomes únicos: `n_distinct(nomb)`.
- Agrupación por provincia e nome para ver total de persoas.
- Top 10 nomes femininos e masculinos en Galicia.
- Código R:

```
nomes %>%  
  filter(sexo == 6) %>%  
  group_by(nomb) %>%  
  summarise(total = sum(cuenta)) %>%  
  arrange(desc(total)) %>%  
  slice_head(n = 10) %>%  
  ggplot(aes(x = reorder(nomb, total), y = total)) +  
  geom_col(fill = "steelblue") +  
  coord_flip() +  
  labs(title = "Top 10 nomes femeninos en Galicia",  
       x = "Nome", y = "Nº persoas")
```

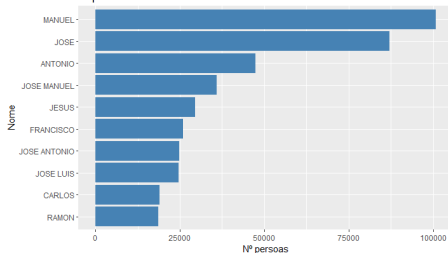
Resumo de nomes en Galicia

- Número de nomes únicos: `n_distinct(nomb)`.
- Agrupación por provincia e nome para ver total de persoas.
- Top 10 nomes femininos e masculinos en Galicia.

Top 10 nomes femininos en Galicia



Top 10 nomes masculinos en Galicia

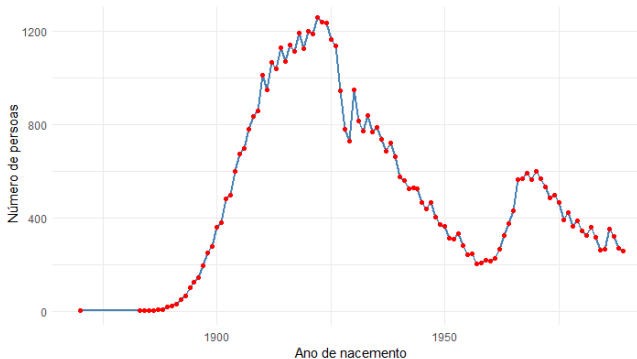


Evolución do nome **MARÍA** ao longo do tempo

```
evolucion_nomes %>%  
  filter(nomb == "MARIA") %>%  
  ggplot(aes(x = ano_nacemento, y = total)) +  
  geom_line(color = "steelblue", size = 1) +  
  geom_point(color = "red") +  
  labs(title = "Evolución do nome 'MARÍA' ao longo do tempo",  
        x = "Ano de nacemento",  
        y = "Número de persoas") +  
  theme_minimal()
```

- Datos agregados por ano de nacemento: `ano_nacemento` = 2000 - idade.
- Visualización con `geom_line()` e puntos co número de persoas por ano.
- Permite identificar tendencias ao longo do tempo.

Evolución do nome **MARÍA** ao longo do tempo



- Datos agregados por ano de nacemento: `ano_nacemento = 2000 - idade`.
- Visualización con `geom_line()` e puntos co número de persoas por ano.
- Permite identificar tendencias ao longo do tempo.

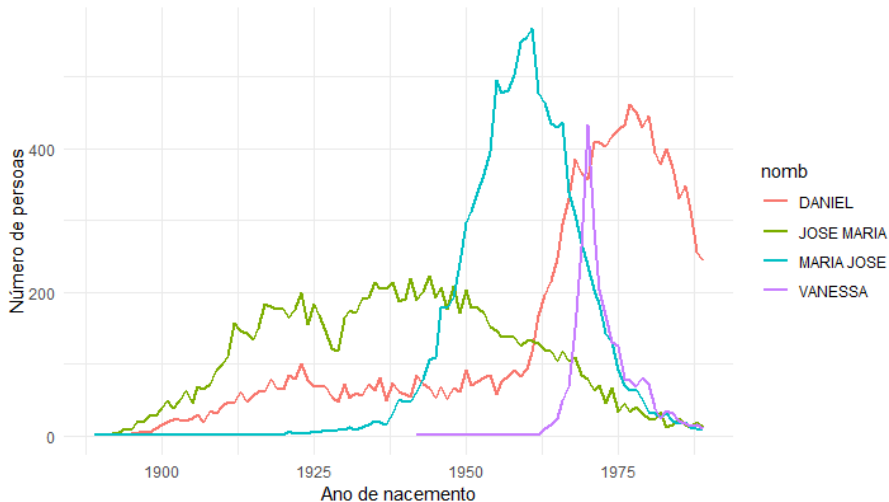
Comparativa de nomes en Galicia

```
evolucion_nomes %>%  
  mutate(decada = floor(ano_nacemento / 10) * 10) %>%  
  group_by(decada, nomb) %>%  
  summarise(total_decada = sum(total, na.rm = TRUE)) %>%  
  arrange(decada, desc(total_decada)) %>%  
  slice_max(total_decada, n = 5) # top 5 nomes por década  
  
evolucion_nomes %>%  
  filter(nomb %in% c("DANIEL", "JOSE MARIA", "MARIA JOSE", "VANESSA")) %>%  
  ggplot(aes(x = ano_nacemento, y = total, color = nomb)) +  
  geom_line(size = 1) +  
  labs(title = "Evolución de nomes en Galicia",  
        x = "Ano de nacemento",  
        y = "Número de persoas") +  
  theme_minimal()
```

- Exemplo con nomes populares: **DANIEL, JOSE MARIA, MARIA JOSE, VANESSA.**
- Permite comparar evolución temporal.
- Visualización con cores distintas para cada nome.

Comparativa de nomes en Galicia

Evolución de nomes en Galicia



Evolución interactiva do nome MARÍA

- Distribución por concello e ano de nacemento.
- Código R:

```
plot_ly(  
  maria_df,  
  x = ~long, y = ~lat,  
  color = ~total,  
  frame = ~ano_nacemento,  
  type = 'scatter',  
  mode = 'markers',  
  marker = list(size = 10, colorscale = 'Viridis',  
    showscale = TRUE),  
  text = ~paste("Concello:", name,  
    "<br>Total:", total,  
    "<br>Ano:", ano_nacemento)  
)
```

Librerías empregadas

- `tidyverse` : fluxo de datos limpo e coherente.
- `sf` , `mapSpain` : manexo de datos espaciais.
- `ggplot2` : visualización estática.
- `mapview` , `plotly` : mapas interactivos.
- `stringr` , `forcats` : limpeza e manipulación de texto.
- `dplyr` : transformación e agregación de datos.

Boas prácticas e conclusións

Boas prácticas aprendidas

- Empregar `dplyr` e `mutate()` para manter o código lexible.
- Combinar datos xeográficos e demográficos con `sf`.
- Validar resultados intermedios antes da visualización.
- Documentar e automatizar o proceso analítico.

Boas prácticas e conclusións

Boas prácticas aprendidas

- Empregar `dplyr` e `mutate()` para manter o código lexible.
- Combinar datos xeográficos e demográficos con `sf`.
- Validar resultados intermedios antes da visualización.
- Documentar e automatizar o proceso analítico.

Resumo e conclusións

- `R` e `tidyverse` permiten análises eficientes e reproducibles.
- A visualización con `ggplot2` é clave para comunicar resultados.
- Novas ferramentas como `Quarto` facilitan a difusión científica.

Recursos recomendados

- **R e Tidyverse:** <https://www.tidyverse.org>
- **R for Data Science:** <https://r4ds.hadley.nz>
- **mapSpain:** <https://ropenspain.github.io/mapSpain>
- **Plotly e Mapview:** para visualización interactiva
- **IGE - Padrón municipal 2000:** <https://www.ige.gal>

Referencias I



Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. Disponible en: <https://r4ds.hadley.nz>



Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the *tidyverse*. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>



Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Disponible en: <https://ggplot2.tidyverse.org>



Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4. Disponible en: <https://CRAN.R-project.org/package=dplyr>

Referencias II



R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Viena, Austria. Disponible en: <https://www.R-project.org>



Instituto Galego de Estatística (IGE). (2000). *Padrón municipal de habitantes 2000*. Santiago de Compostela: IGE. Disponible en: <https://www.ige.gal>



Hernangómez, D. (2024). *mapSpain: Administrative Boundaries of Spain*. R package version 0.5.0. Disponible en: <https://CRAN.R-project.org/package=mapSpain>



Pebesma, E. (2023). *Simple Features for R: Standardized Support for Spatial Vector Data*. *The R Journal*, 10(1), 439–446. <https://doi.org/10.32614/RJ-2018-009>



Sievert, C. (2022). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC. Disponible en: <https://plotly-r.com>

O **tidyverse** non é só un conxunto
de paquetes...
é unha forma de pensar con datos.

TRABALLAR CON DATOS DE XEITO LIMPO E EFICIENTE CO UNIVERSO *tidyverse*

María José Ginzo Villamayor

Universidade de Santiago de Compostela
Centro de Investigación e Tecnoloxía Matemática de Galicia

16 de outubro de 2025



DEPARTAMENTO DE ESTATÍSTICA,
ANÁLISE MATEMÁTICA E OPTIMIZACIÓN



CENTRO DE INVESTIGACIÓN
E TECNOLOXÍA MATEMÁTICA
DE GALICIA

