Limpieza de datos en R - Temas adicionales

Joins o juntar tablas

- Los `joins` son operaciones que permiten combinar dos o más tablas en base a una o más columnas comunes.
- Llamamos a las columnas comunes *keys*, *join jeys* o identificadores.
- Un join se puede visualizar como una función `BUSCARV` de Excel, donde se busca un valor en una tabla y se lo añade a otra tabla.
- Existen varios tipos de joins, los más comunes son `inner`, `left`, `right` y `full`.

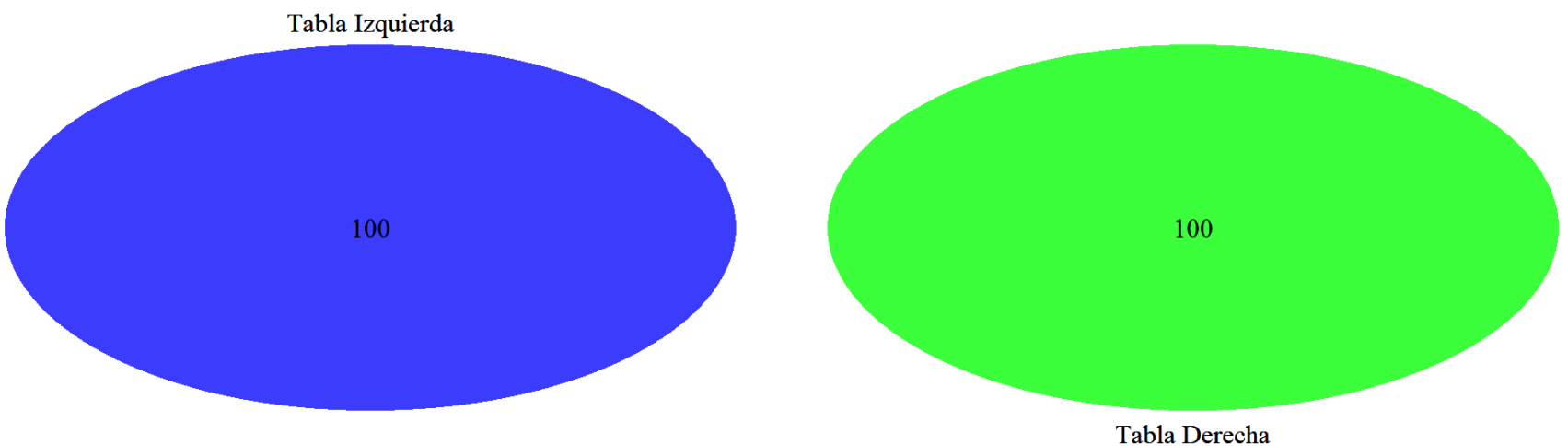
¿Por qué usar joins?

- A veces, requiero datos de diferentes fuentes para responder una pregunta de investigación.
- Ej. datos de encuesta de empleo ENEMDU deben unirse con datos de nutrición ENSANUT.
- Los identificadores comunes se suelen definir a niveles de agregación (ej. provincia, canton, parroquia, etc.)
- Algunas bases de datos **exigen** que se unan tablas para poder analizarlas. Ej. resultados electorales CNE.

Terminología de los joins

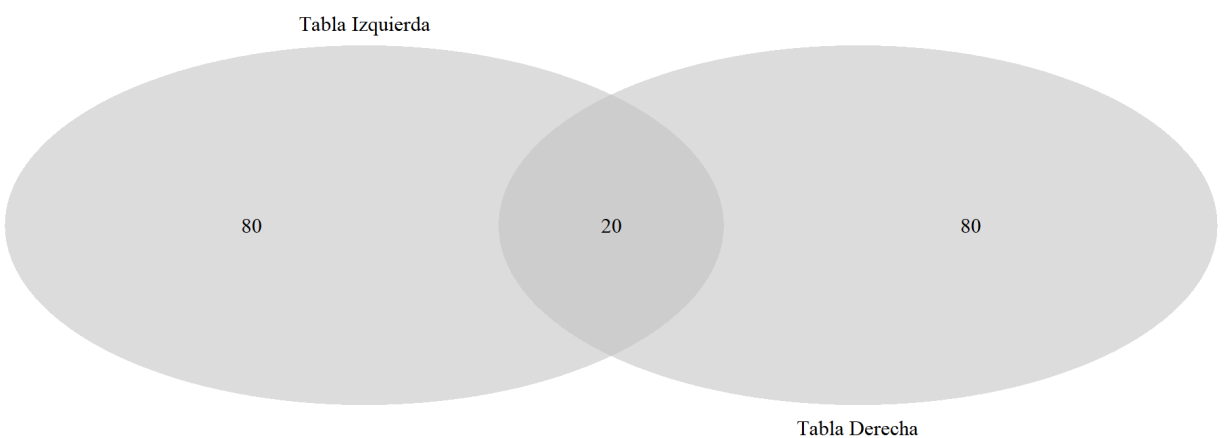
- Surgen de la terminología de bases de datos relacionales (SQL)
- Se define una tabla principal o `left` y una tabla secundaria o `right` (tablas de la izquierda y derecha)
- Se visualiza más facilmente como diagramas de Venn

Ejemplo: Dos tablas con 100 filas cada una



Inner Join - `inner_join()`

- Solamente se quedan las filas que tienen valores en comun en ambas tablas (definidos por el `key`)
- Abajo, solo conservamos las 20 filas que tienen valores en común en ambas tablas

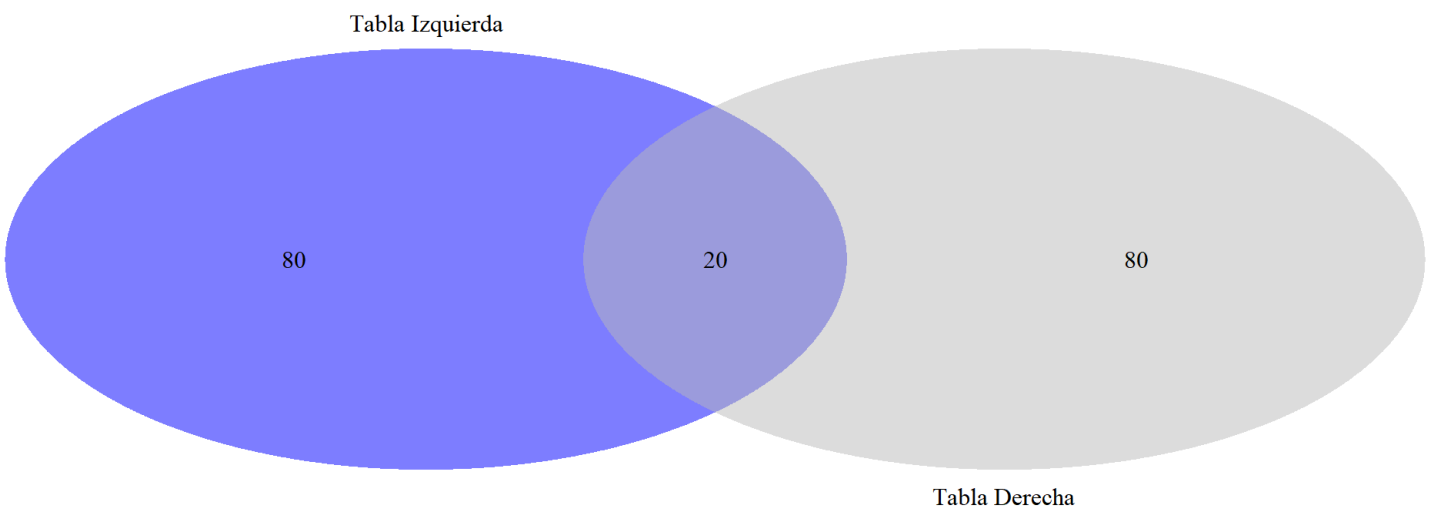


Inner Join - implementación en código

```r
library(dplyr)

# Crear bases de datos

df1 <- data.frame(key = c(1, 2, 3, 4, 5),
                  value = c("A", "B", "C", "D", "E"))

df2 <- data.frame(key = c(3, 4, 5, 6, 7),
                  value = c("C", "D", "E", "F", "G"))

# Inner Join

df_resultado <- inner_join(df1, df2, by = "key")

df_resultado
```

Left Join - `left_join()`

- Se quedan todas las filas de la tabla izquierda y las filas de la tabla derecha que tengan valores en común
- Conservamos 80 filas de la tabla izquierda.



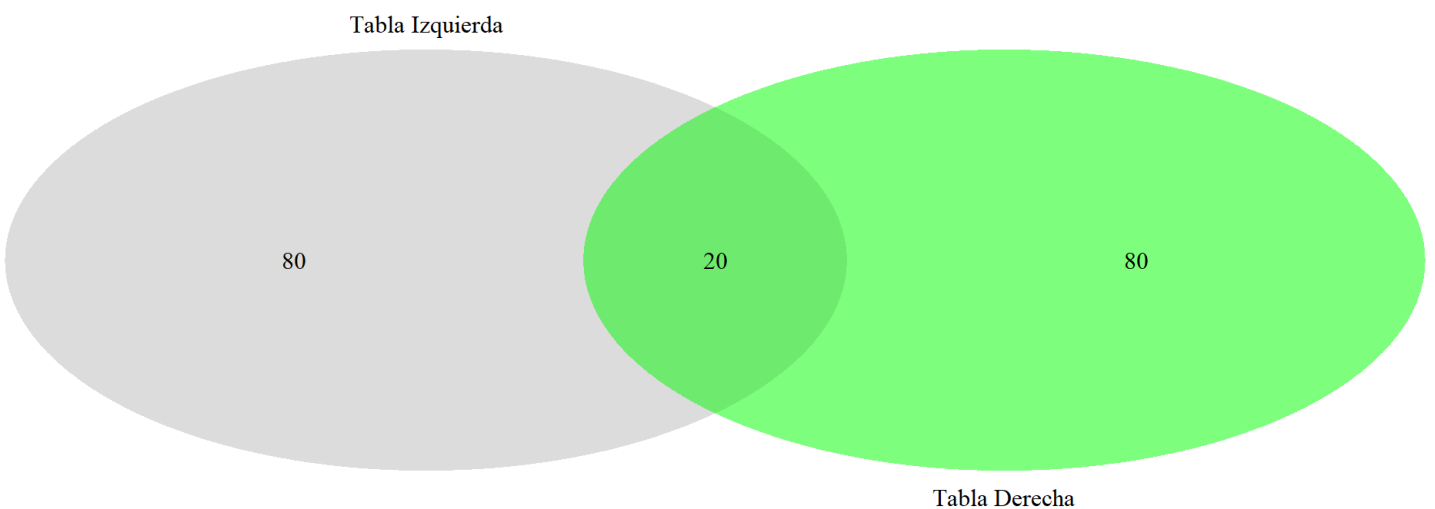
Left Join - implementación en código

```r
# Left Join

df_resultado <- left_join(df1, df2, by = "key")

df_resultado
```

```
  key value.x value.y
1   1       A    <NA>
2   2       B    <NA>
3   3       C       C
4   4       D       D
5   5       E       E
```

Right Join - `right_join()`

- Se quedan todas las filas de la tabla derecha y las filas de la tabla izquierda que tengan valores en común
- Conservamos 80 filas de la tabla derecha.



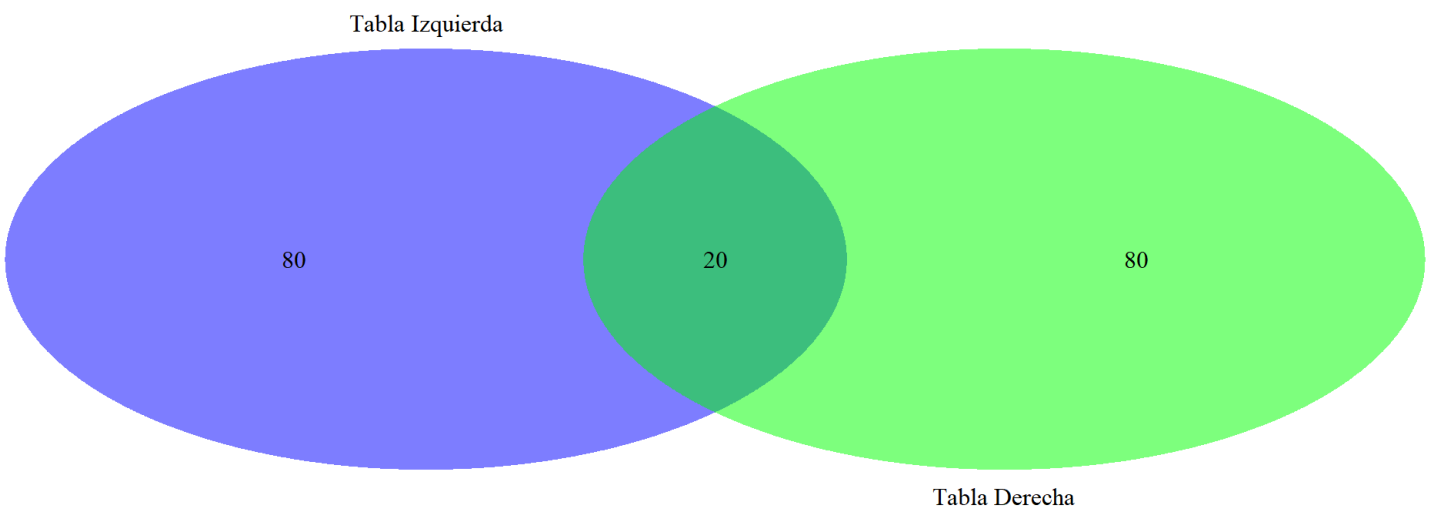
Right Join - implementación en código

```r
# Right Join

df_resultado <- right_join(df1, df2, by = "key")

df_resultado
```

```
  key value.x value.y
1   3       C       C
2   4       D       D
3   5       E       E
4   6    <NA>       F
5   7    <NA>       G
```

Full Join - `full_join()`

- Se quedan todas las filas de ambas tablas, conservando los valores en común y los valores únicos de cada tabla.
- Conservamos 100 filas.



UNION de tablas

- Las uniones de tablas son operaciones que permiten juntar dos tablas en una sola tabla, sin importar si tienen columnas en común.
- Se pueden pensar como uniones horizontales, mientras que los `JOINS` son verticales, a partir de un identificador.
- En R, se pueden hacer con la función `bind_rows()` de la librería `dplyr`.
- En un contexto investigativo, se definen las `UNION` con encuestas o bases de datos que tienen las mismas columnas a lo largo de diferentes archivos.
    - Ej. encuestas de empleo ENEMDU 2015, 2016, 2017, etc.

`bind_rows()` - implementación en código

```
   key value year
1    1     A 2015
2    2     B 2015
3    3     C 2015
4    4     D 2015
5    5     E 2015
6    6     F 2016
7    7     G 2016
8    8     H 2016
9    9     I 2016
10  10     J 2016
```