Visualización de datos: PEC 2.

Máster Universitario en Ciencia de Datos. UOC

Gabriel Nicolás Rivadeneira Gómez

2025-11-02

Índice

1. Técnica avanzada. Gráfico matricial.

1.1. Carga de datos.

Primero cargo los datos preprocesados anteriormente. Estos son una versión larga, con 6 columnas

PROVINCIA, YEAR, SECTOR, PARADOS, POBLACION, PARADOS POR CADA MIL HABs

```
csv_path<-"datos_paro_andalucia.csv"
 df<-read.csv(csv_path,sep=";",header=TRUE, encoding = "latin1")
```

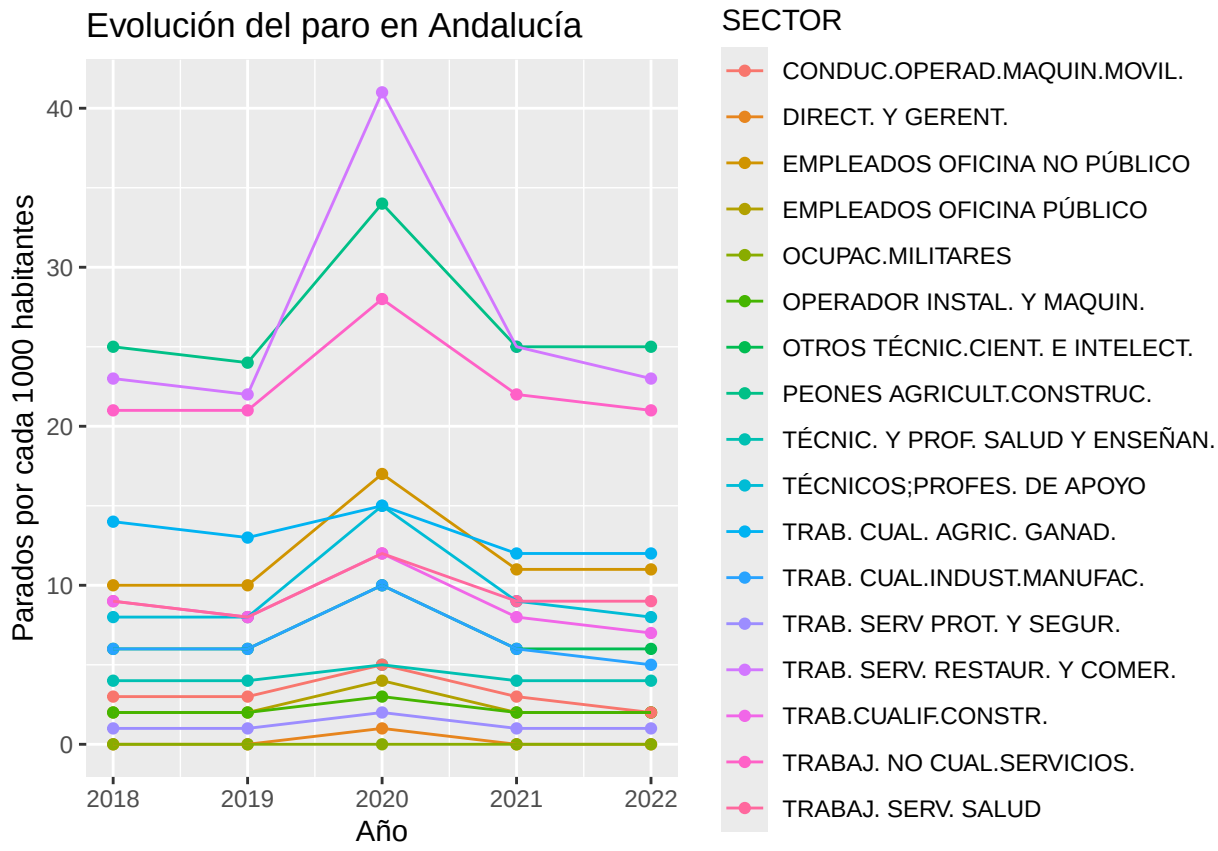
```
colnames(df)
```

```
## [1] "PROVINCIA"                "YEAR"
## [3] "SECTOR"                   "PARADOS"
## [5] "POBLACION"                "PARADOS.POR.CADA.MIL.HABs"
```

1.2. Revisión holística de lo datos.

A continuación comparo todas las lineas para buscar patrones relevantes. Esta revisión viene motivada por la idea de que los sectores más afectados por la pandemia son los relacionados con la hostelería, comercio y turismo.

```
ggplot(df, aes(x = YEAR, y = PARADOS.POR.CADA.MIL.HABs, colour = SECTOR)) +
  geom_point() +
  geom_line() +
  theme(legend.position = "right")+
  labs(title = "Evolución del paro en Andalucía",
       x = "Año",
       y = "Parados por cada 1000 habitantes")
```



Se aprecia que los sectores más afectados son los relacionados con el turismo y restauración, además de otros sectores como la industria o construcción, como se ha mencionado en el mapa coroplático de la técnica básica.

1.3. Generación de la visualización con técnica avanzada.

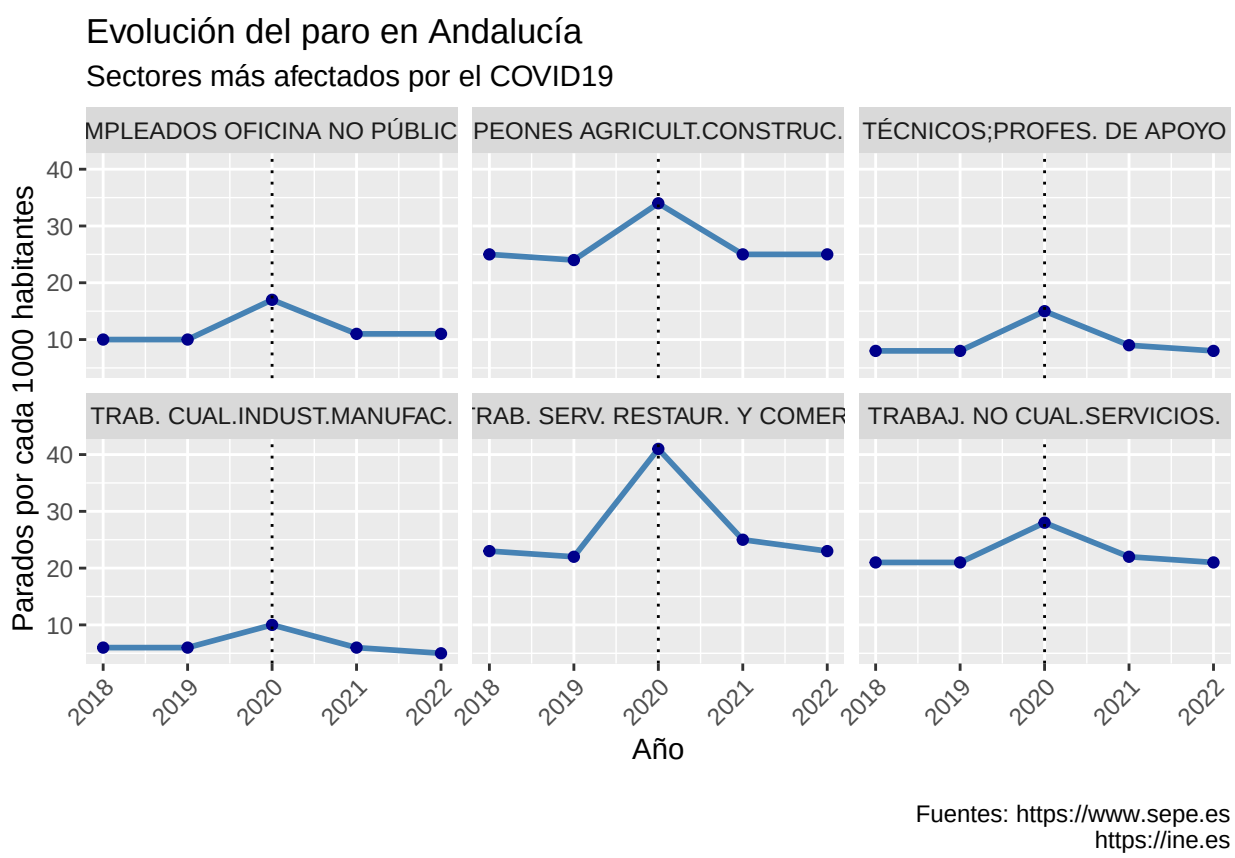
Ahora genero la visualización matricial siguiendo el ejemplo del recurso **Introducción a ggplot2 y ggmap**, apt 1.2.2 Estéticas.

Para ello primero genero un dataframe con los sectores de interés. En este caso: ["TRAB. SERV. RESTAUR. Y COMER.", "TRABAJ. NO CUAL.SERVICIOS.", "PEONES AGRICULT.CONSTRUC.", "EMPLEADOS OFICINA NO PÚBLICO", "TÉCNICOS;PROFES. DE APOYO", "TRAB. CUAL.INDUST.MANUFAC."]

```
#Extraigo los sectores que me interesan
sectores_objetivo <- c("TRAB. SERV. RESTAUR. Y COMER.",
                       "TRABAJ. NO CUAL.SERVICIOS.",
                       "PEONES AGRICULT.CONSTRUC.",
                       "EMPLEADOS OFICINA NO PÚBLICO",
                       "TÉCNICOS;PROFES. DE APOYO",
                       "TRAB. CUAL.INDUST.MANUFAC."
                       )
#Me quedo con los datos relacionados del primer dataframe
df_filtrado <- df[df$SECTOR %in% sectores_objetivo, ]

#Pinto la matriz de representaciones.
ggplot(df_filtrado, aes(x = YEAR, y = PARADOS.POR.CADA.MIL.HABs)) +
  geom_line(color = "steelblue", linewidth = 1) +
  geom_point(color = "darkblue") +
```

```
  geom_vline(xintercept = 2020, color="black", linewidth=0.5, linetype="dotted") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~ SECTOR, scales = "fixed") +
  labs(title = "Evolución del paro en Andalucía",
       subtitle = "Sectores más afectados por el COVID19",
       x = "Año",
       y = "Parados por cada 1000 habitantes",
       caption = "
       Fuentes: https://www.sepe.es\nhttps://ine.es")
```




```
#Guardo la figura
ggsave('tecnica_avanzada_pec2.png')
```

```
## Saving 6.5 x 4.5 in image
```

He decidido rotar las etiquetas del eje X 45º para mejorar su lectura. He optado por usar la opcion *scales = "fixed"* para proporcionar una visión clara de los sectores con más paro en términos de ratio absoluto. La línea punteada marca el año del covid, que se propagó por España a comienzos del 2020.

1.4. Representación completa de los datos.

La misma representación con todos los datos es la siguiente:

Evolución del paro por sectores en Andalucía.

Influencia del covid19

Fuentes: https://www.sepe.es
https://ine.es

2. Técnica específica. Gráfico de enjambre.

2.1. Carga de datos.

Primero cargo los datos necesarios. Estos son una versión larga, con 7 columnas:

PROVINCIA, TOTAL, AÑO, POBLACION, RATIO, EDAD, PARADOS

- Provincia: El campo que indica el nombre de la provincia.
- Total: Es el valor de parados totales en todas las edades de la provincia, en valor absoluto.
- Año: El año de estudio. En este caso el año 2020.
- Ratio: El número de parados en ese grupo de edad por cada mil habitantes de esa provincia.
- Edad: Indica el grupo de edad. Hay 10 grupos de edad; [DE 16-19 AÑOS,DE 20-24 AÑOS,DE 25-29 AÑOS,DE 30-34 AÑOS,DE 35-39 AÑOS,DE 40-44 AÑOS,DE 45-49 AÑOS,DE 50-54 AÑOS,DE 55-59 AÑOS,MAYOR DE 59 AÑOS]
- Parados: El número de parados para el grupo de edad.

```
csv_path<-"paro_provincias_andalucia_enjambre.csv"
 df_enjambre<-read.csv(csv_path,sep=";",header=TRUE, encoding = "latin1")
```

Compruebo que se han cargado las columnas descritas.

```
colnames(df_enjambre)
```

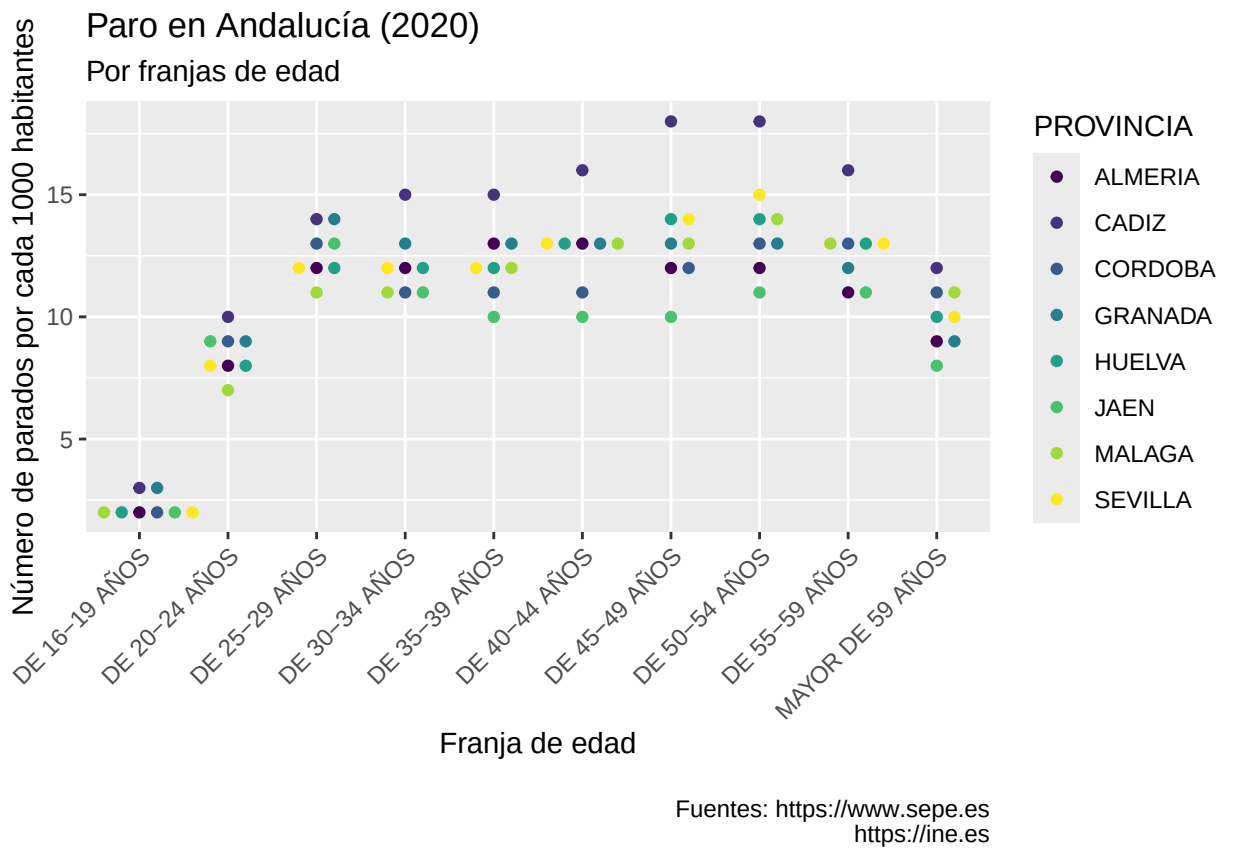
```
## [1] "PROVINCIA" "TOTAL"     "AÑO"       "POBLACION" "RATIO"     "EDAD"
## [7] "PARADOS"
```

2.2. Generación de la visualización con técnica específica.

Ahora genero la visualización usando de guía el apartado anterior. En este caso la capa que voy a usar es beeswarm.

La figura representa el ratio de paro (por cada mil habitantes) por franja de edad.

```
#Pinto el enjambre.
ggplot(df_enjambre, aes(x = EDAD, y = RATIO, colour = PROVINCIA)) +
  geom_beeswarm(priority = "density", cex = 2, alpha = 1) +
  scale_color_viridis_d()+
  labs(title = "Paro en Andalucía (2020)",
       subtitle = "Por franjas de edad",
       x = "Franja de edad", y = "Número de parados por cada 1000 habitantes",
       caption = "
       Fuentes: https://www.sepe.es\nhttps://ine.es")+
  #Como se indica en la guía ese pueden modificar aspectos del tema.
  #En este caso modifico el angulo de las etiquetas.
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



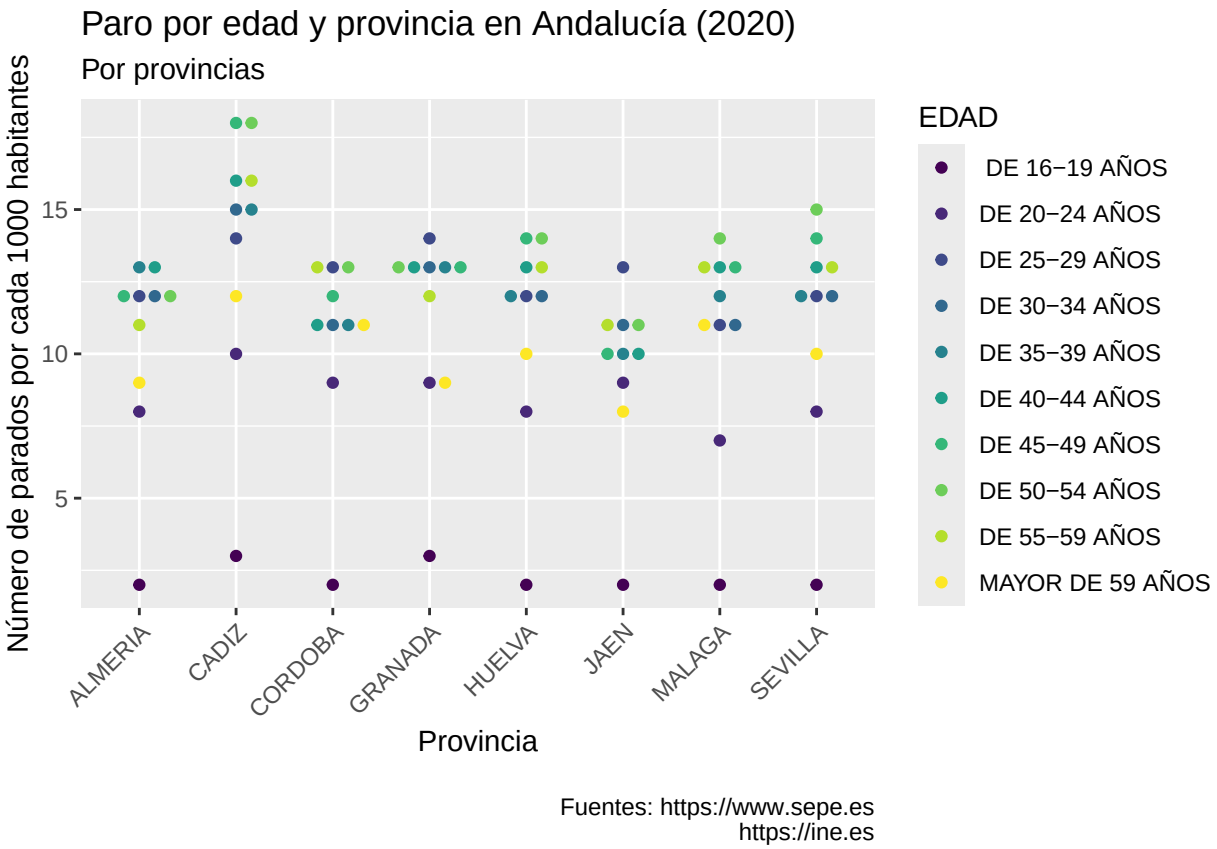

```
ggsave("beeswarm_paro_andalucia_edad.png")
```

```
## Saving 6.5 x 4.5 in image
```

Cada punto de la figura anterior representa el ratio de paro de una provincia, agrupado por rango de edad. Se puede observar que la franja de edad de 16 a 19 años es la que menor ratio de paro presenta. Esto puede tener explicación en que los datos tomados, como ya se ha mencionado, corresponden a demandas de empleo procedentes de personas ya paradas, es decir, que ya han tenido antes un trabajo, con lo cual en este grupo de edad existirá poco paro en el sentido estricto porque no se contabilizan perfiles que no hayan tenido un trabajo anterior, lo que es común en personas de estas edades. En linea con lo anteiror, el paro parece equiparase en todas las franjas.

Cabe destacar que para Cádiz, los puntos muestran que en todas las franjas de edad a partir de 30 años, el paro es superior que para el resto de provincias.

Para ahondar en este detalle, he generado la misma visualización pero representando el ratio de paro en función de la provincia.




```
## Saving 6.5 x 4.5 in image
```

En la visualización ahora cada punto representa el ratio de paro en cada provincia, y los colores indican el rango de edad.

Se puede observar cómo efectivamente Cádiz tiene un ratio de paro mayor desde los 30 hasta los 59 años.

De nuevo podemos ver que las personas más jóvenes tienen muy poco paro, en general, en todas las provincias de Andalucía por el motivo explicado antes.

3. Referencias.

Para la lectura de fichero de datos csv, la generación de las diferentes figuras y generación del fichero resultado he consultado las siguientes fuentes.

- **Gil Bellosta, C. J.** *Introducción a ggplot2 y ggmap*. Universitat Oberta de Catalunya. PID_00235524.
- **Rivadeneira Gómez, G. N. (2025)**. *Actividad 1*. Trabajo presentado en la asignatura Análisis estadístico. Máster en Ciencia de Datos, Universidad Oberta de Catalunya.
- **Rivadeneira Gómez, G. N. (2025)**. *Actividad 2*. Trabajo presentado en la asignatura Análisis estadístico. Máster en Ciencia de Datos, Universidad Oberta de Catalunya.
- *https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf*
- Datos de demandas de empleo: *https://www.sepe.es/HomeSepe/que-es-el-sepe/estadisticas/empleo/estadisticas-nuevas*
- Datos de población: *https://ine.es*