MDPI

Article

Estudio del gasto y duración media de los viajes de los turistas extranjeros en distintas comunidades autónomas

Alejandro León Líndez[1], Adrian Lizzadro Pla[2], Marta Medina Muñiz[3]

[1] Máster en Ciencia de Datos; alelin@alumni.uv.es
[2] Máster en Ciencia de Datos; alizpla@alumni.uv.es
[3] Máster en Ciencia de Datos; memuiz@alumni.uv.es
* Correspondence: email@email.com; Tel.: +XX-000-00-0000.

Simple Summary: Resumen del trabajo

Abstract: Estudiar la evolución del gasto total, el gasto medio por persona y el gasto medio diario en el periodo de 2016 a 2023 de turistas con diversos países de residencia en distintas comunidades autónomas. Estudio de la duración media de dichos viajes y su relación con el gasto por persona.

Keywords: keyword 1; keyword 2; keyword 3 (list three to ten pertinent keywords specific to the article, yet reasonably common within the subject discipline.).

1. Introducción

2. Carga de librerías e importación del fichero

Antes de comenzar, eliminamos todas las variables guardadas.

```
rm(list = ls())  # Borrado de todas las variables
```

```
library(readr)  # Librería para importación de datos
library(dplyr)  # Librería para arreglo de datos
library(tidyr)  # Librería para arreglo de datos
library(ggplot2)  # Librería para gráficas
library(sf)  # Librería para generar el mapa geográfico
library(giscoR)  # Librería para generar el mapa geográfico
```

Importamos los datos

```
gastos <- read_delim("data/Gasto_turistas_internacionales_segun_comunidad_paisresi
    delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

3. Preparación de los datos

3.1. Tranformación a tidydata

Antes de comenzar con el preprocesamiento de los datos, observamos el conjunto de datos importado en la Tabla 1.

Citation: Estudio del gasto y duración media de los viajes de los turistas extranjeros en distintas comunidades autónomas. *Data* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

Table 1. 10 primeras filas de los datos importados

País de residencia	Total Nacional y CCAA	Tipo de dato	Gastos y duración media de los viajes	Periodo	Total
Total	Total	Dato base	Gasto total	2023	108.789,41
Total	Total	Dato base	Gasto total	2022	87.138,19
Total	Total	Dato base	Gasto total	2021	34.903,37
Total	Total	Dato base	Gasto total	2020	19.786,78
Total	Total	Dato base	Gasto total	2019	91.911,97
Total	Total	Dato base	Gasto total	2018	89.750,75
Total	Total	Dato base	Gasto total	2017	87.003,93
Total	Total	Dato base	Gasto total	2016	77.415,54
Total	Total	Dato base	Gasto medio por persona	2023	1.277
Total	Total	Dato base	Gasto medio por persona	2022	1.216

Transformamos este conjunto de datos a un conjunto tidy, donde cada variable se encuentre en una columna y eliminamos filas que no vamos a emplear en el estudio (tasa de variación) y columnas redudantes o que contienen información irrelevante.

De esta manera, obtenemos 4 variables a partir de la columna Gastos y duración media de los viajes: el gasto total, el gasto medio por persona, el gasto medio diario por persona y la duración media de los viajes; cuyos valores son los correspondientes a la columna Total.

Por otro lado, en la columna Tipo de dato contamos con dos valores: Dato base y Tasa de variación anual. Vamos a trabajar únicamente con los datos base, por lo que eliminamos las filas correspondientes a Tasa de variación anual y eliminamos la columna Tipo de dato, que ahora aporta información redundante.

3.2. Cambios de nombres de las columnas

Renombramos las columnas de manera apropiada (sin espacios y con nombres representativos de las variables). Los nombres de las variables son los siguientes: Pais, CCAA, Periodo, Gasto_total, Gasto_medio_persona, Gasto_medio_diario_persona y Duracion_media.

3.3. Transformación de clases

Todas las variables han sido importadas como tipo carácter. Transformamos las variables Gasto_total, Gasto_medio_persona, Gasto_medio_diario_persona y Duracion_media a numérico, donde previamente transformamos la cadena de carácteres a una que sea interpretable como número para poder aplicar la función as.numeric() correctamente (eliminando el punto de miles y sustituyendo la coma decimal por un punto decimal).

Comprobamos si al realizar la transformación se han introducido datos NA y a continuación, transformamos a factor las variables Pais y CCAA.

Quitamos algunas filas que contienen datos que consisten en la suma total de los datos de la columna CCA (más adelante eliminaremos también los valores correspondientes a Total de la variable Pais). Sin embargo, recogeremos estos valores en datasets datos_totales y datos_total_por_residencia para hacer uso de ellos y obtener información relevante. Cambiamos los nombres de los niveles del factor CCAA a los siguientes: “Andalucía”, “Illes Balears”,“Canarias”, “Cataluña”,“Comunitat Valenciana”, “Comunidad de Madrid”, “Otras CCAA”.

3.4. Instance engineering

Vamos a trabajar con valores de las filas para obtener un dataset más apropiado al estudio que deseamos realizar. En primer lugar, queremos deshacernos del nivel “Total” en la variable Pais y transformarla en otro llamado “Otros” en que en el resto de variables contenga la información referente al resto de paises distintos de los cuáles poseemos datos concretos.

3.4.1. Instance engineering de Gasto_total

Para la variable Gasto_total, las filas correspondientes a "Otros" de la variable Pais y las filas correspondientes a "Total" se relacionan de la siguiente manera con $pais \in \{Alemania, Francia, Italia, Pases\ Nrdicos, Reino\ Unido, Otros\}$:

$$\text{Gasto total}_{\text{Total}} = \sum_{pais} \text{Gasto total}_{pais}$$

Luego el gasto total de "Otros" es el gasto de los valores "Total" de Pais menos el gasto de cada uno de los otros 5 paises disponibles por separado.

3.4.2. Instance engineering de Gasto_medio_persona, Gasto_medio_diario_persona y Duracion_media

Como en estas variables se trata de una media, no podemos emplear el método usado para Gasto_total. En su lugar consideramos una media ponderada. Si consideramos que el número de paises totales es 194 tenemos que para $pais \in \{Alemania, Francia, Italia, Pases\ Nrdicos, R$

$$\overline{Media\ Total} = \frac{5}{194} * \sum_{pais} Valor\ medio_{pais} + \frac{194-5}{194} * Valor\ Medio_{otros}$$

De esta manera, concemos el valor de $\overline{Media\ Total}$ y de $Valor\ medio_{pais}$ y podemos calcular el valor de los valores medios para las filas "Otros" de la variable Pais, otorgándoles un mayor peso de manera proporcional al número de países considerados en esta categoría.

4. Resumen de los datos

Una vez hemos concluido el pre-procesamiento de los datos, observamos el resultado en la Tabla 2.

Table 2. 10 primeras filas del conjunto de datos procesados

	Pais	CCAA	Periodo	Gasto_total	Gasto_medio_persona	Gasto_medio_diario_persona	Duracion_media
1	Alemania	Andalucía	2016	1121.33	1132	102	11.14
2	Alemania	Andalucía	2017	1258.89	1123	105	10.70
3	Alemania	Andalucía	2018	1256.69	1165	114	10.18
4	Alemania	Andalucía	2019	1168.38	1048	117	8.96
5	Alemania	Andalucía	2020	255.96	1056	102	10.34
6	Alemania	Andalucía	2021	464.01	1040	100	10.39
7	Alemania	Andalucía	2022	1045.65	1211	113	10.69
8	Alemania	Andalucía	2023	1314.95	1267	130	9.71
17	Alemania	Illes Balears	2016	4436.99	961	129	7.44
18	Alemania	Illes Balears	2017	4890.66	1013	133	7.63

Vamos un resumen de los datos y de las variables disponibles. En la Tabla 3 se ha realizado un pequeño codebook con las variables del dataset.

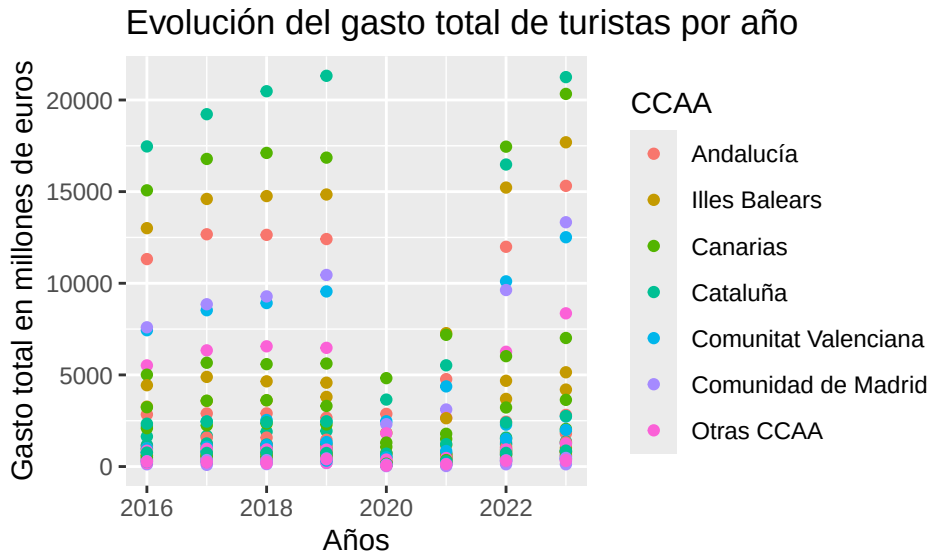
Table 3. Descripción de las variables

Nombre variable	Unidad	Valores
Pais	-	Alemania, Italia, Países Nórdicos, Francia, Reino Unido, Otros
CCAA	-	Andalucía, Illes Balears, Canarias, Cataluña, Comunitat Valenciana, Comunidad de Madrid, Otras CCAA
Periodo	Año	2016-2023
Gasto_total	Millones de euros	Numérico
Gasto_medio_persona	Euros	Numérico
Gasto_medio_diario_persona	Euros	Numérico
Duración_media	Días	Numérico

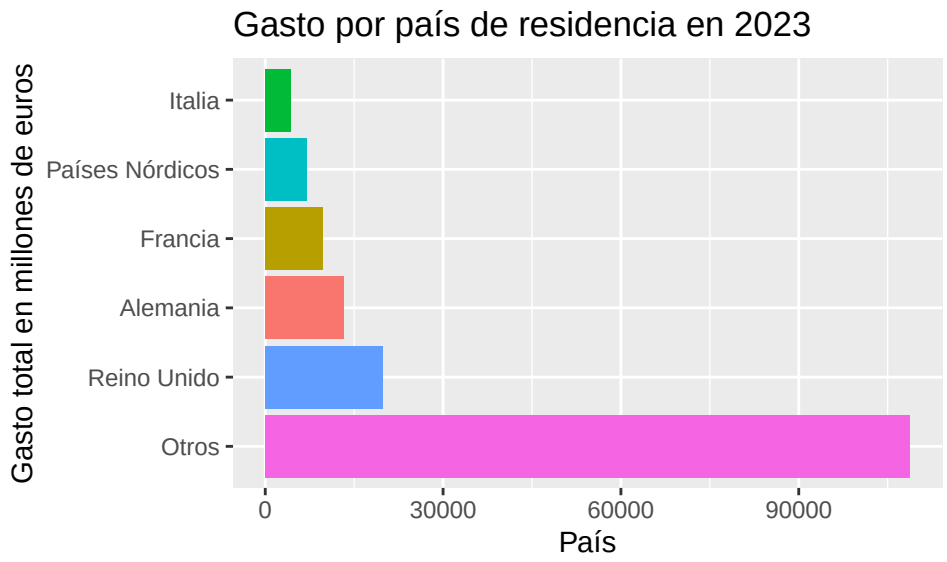
```
# Hacer el resumen con dplyr no summary. Tengo que hacerlo
```

5. Prueba de ggplot

```
# Visualización de los datos e interpretación de los
# posibles patrones

# Visualización previa para tener una idea de ciertas
# variables de los datos
ggplot(datos, aes(x = Periodo, y = Gasto_total, color = CCAA,
    group = 1)) + geom_point() + labs(title = "Evolución del gasto total de turist
    x = "Años", y = "Gasto total en millones de euros")
```

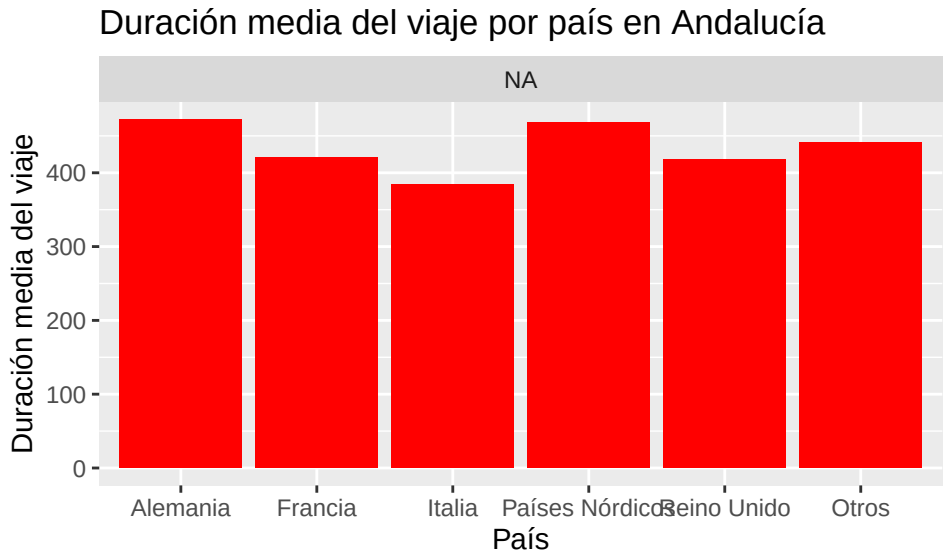



```
# Nos centramos en un año en concreto y visualizamos el
# gasto total por país en ese año
ggplot(datos[datos$Periodo == 2023, ], aes(x = reorder(Pais,
    -Gasto_total), y = Gasto_total)) + geom_bar(stat = "identity",
    aes(fill = Pais)) + coord_flip() + labs(title = "Gasto por país de residencia
    x = "Gasto total en millones de euros", y = "País") + theme(legend.position =
```



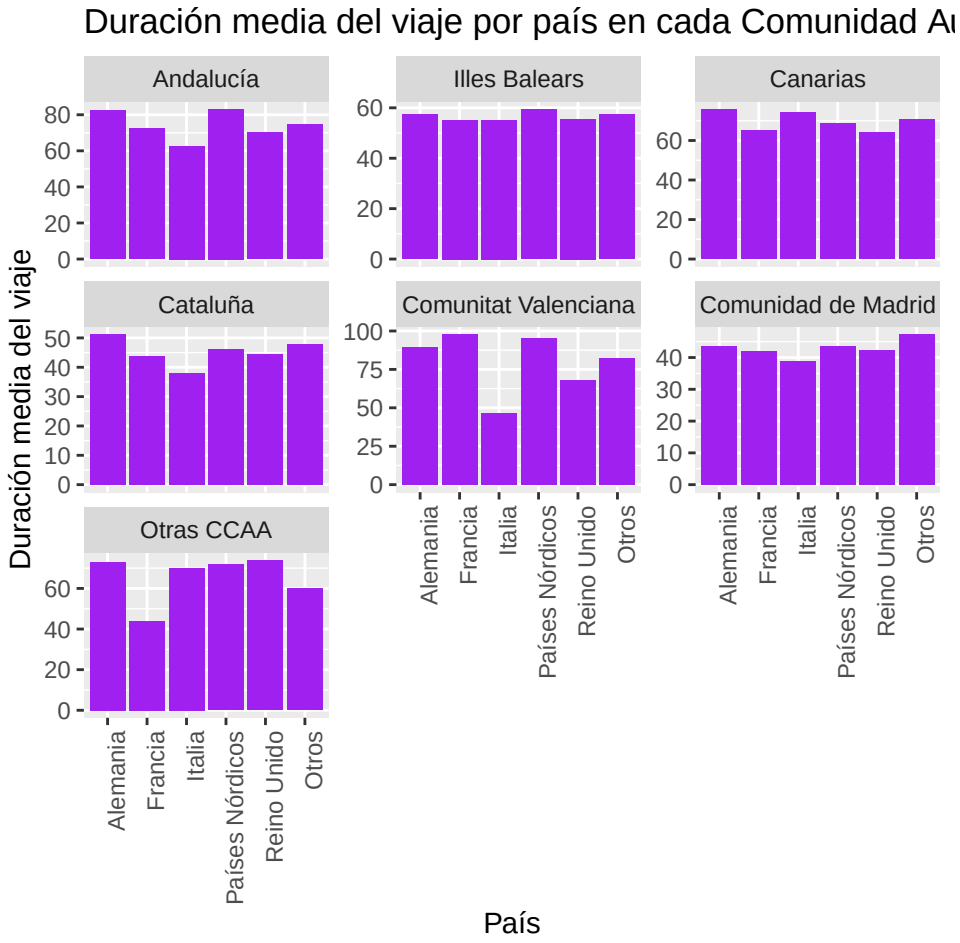
72

```r
# Para analizar únicamente una variable por país en una
# sola Comunidad Autónoma.
ggplot(datos, aes(x = Pais, y = Duracion_media)) + geom_bar(stat = "identity",
    fill = "red") + facet_grid(. ~ CCAA["i"], scales = "free") +
    labs(title = "Duración media del viaje por país en Andalucía",
        x = "País", y = "Duración media del viaje")
```



73

```r
# Para todas las Comunidades Autónomas por separado
ggplot(datos, aes(x = Pais, y = Duracion_media)) + geom_bar(stat = "identity",
    fill = "purple") + facet_wrap(~CCAA, scales = "free_y") +
    labs(title = "Duración media del viaje por país en cada Comunidad Autónoma",
        x = "País", y = "Duración media del viaje") + theme(axis.text.x = element_
    hjust = 1))
```



6. Probamos a analizar el efecto del COVID en el gasto de los turistas, así como los países que más reducieron su gasto debido a la pandemia.

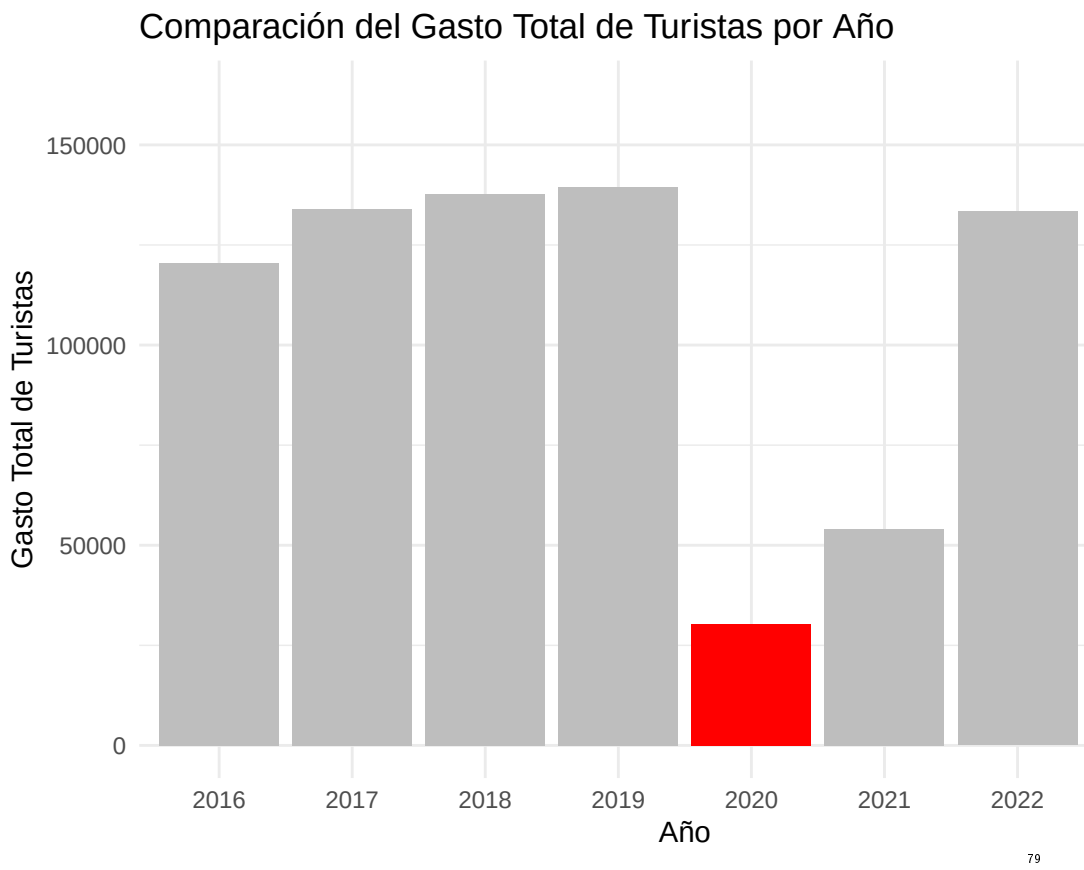
```
library(tidyr)

covid <- datos[datos$Periodo == 2020, ]
no_covid <- datos[datos$Periodo != 2020, ]
media_sin_covid <- mean(no_covid$Gasto_total[])
media_covid <- mean(covid$Gasto_total)
cat("La media del gasto total durante el COVID es de", media_covid,
    "millones de euros\n")
```

```
## La media del gasto total durante el COVID es de 719.3212 millones de euros
```

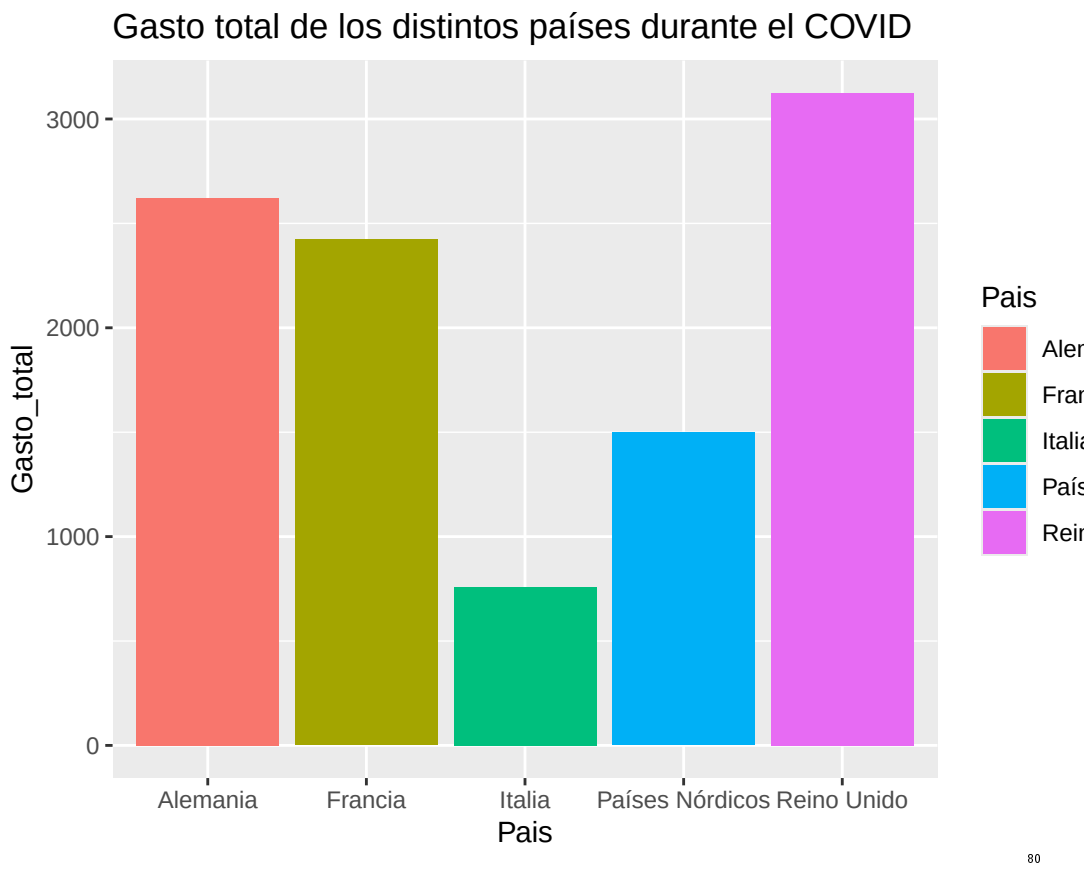
```
cat("Mientras que sin COVID es de", media_sin_covid, "millones de euros\n")
```

```
## Mientras que sin COVID es de 2999.059 millones de euros
```

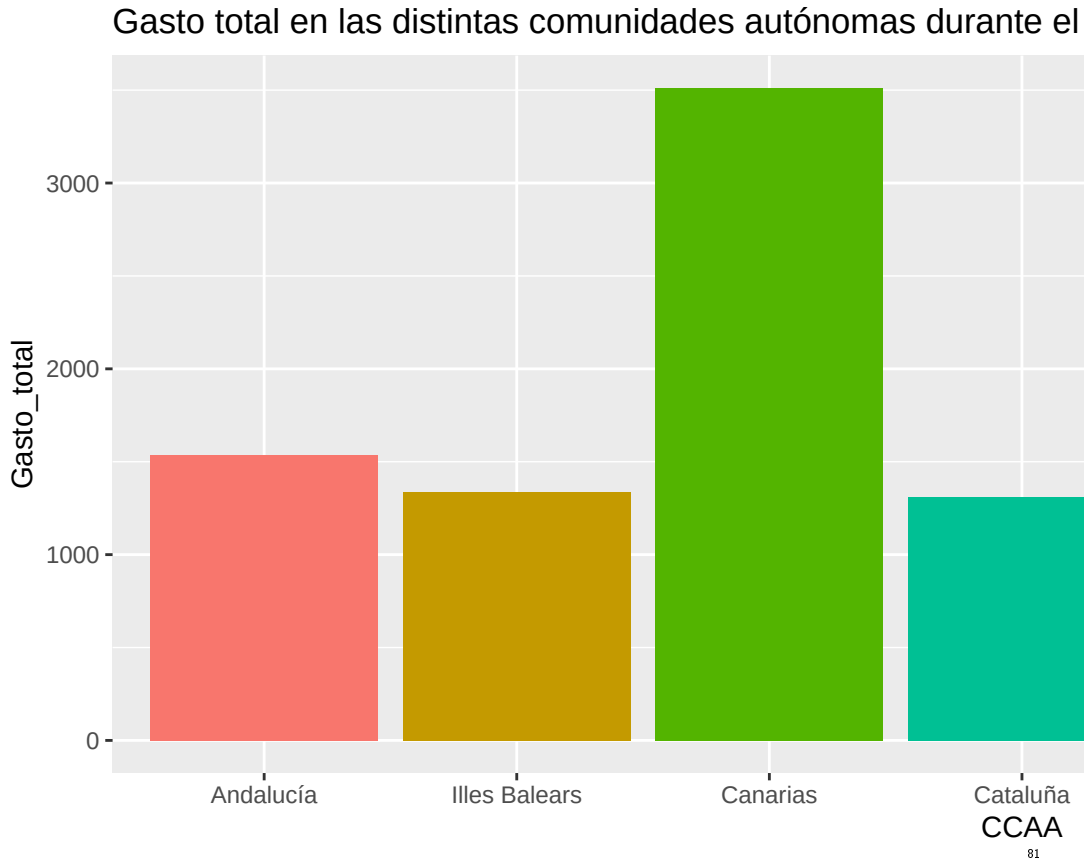
```
datos %>%
    ggplot(aes(x = factor(Periodo), y = Gasto_total, fill = (Periodo ==
        2020))) + geom_bar(stat = "identity") + scale_fill_manual(values = c("grey
    "red")) + labs(x = "Año", y = "Gasto Total de Turistas",
    title = "Comparación del Gasto Total de Turistas por Año") +
    theme_minimal() + theme(legend.position = "none")
```




```
covid_sin_otros <- covid[covid$Pais != "Otros", ]  # Elimino los datos correspondi
covid_sin_otros %>%
    ggplot(aes(x = Pais, y = Gasto_total, fill = Pais)) + geom_bar(stat = "identit
    labs(title = "Gasto total de los distintos países durante el COVID")  # Reino
```




```
# ¿En qué comunidad autónoma se viajó más durante el COVID?
covid_sin_otros %>%
    ggplot(aes(x = CCAA, y = Gasto_total, fill = CCAA)) + geom_bar(stat = "identit
    labs(title = "Gasto total en las distintas comunidades autónomas durante el CO
```

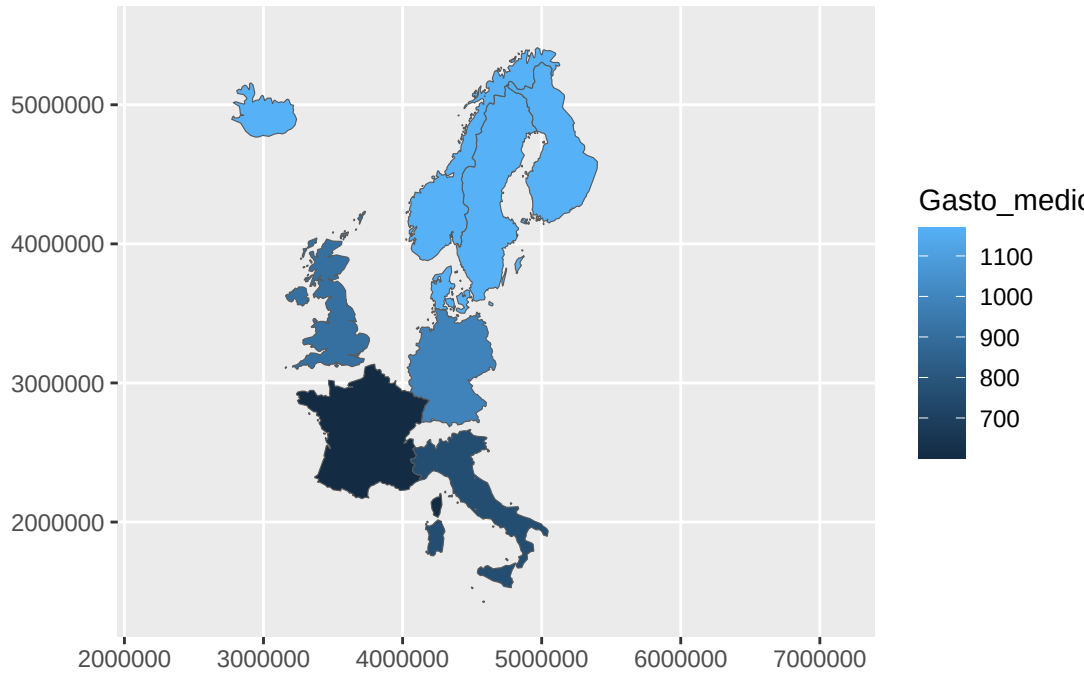



```
# En Canarias el gasto de turistas fue mayor que en el
# resto de Comunidades Autónomas durante el COVID, ¿menor
# regulación?
```

Filtraremos los datos totales por el Periodo y recogeremos solo la variable que nos sirve para representar la magnitud de valores en el mapa de visualización, en este caso el gasto medio por persona

```
gastos_medio_residencia <- datos_total_por_residencia %>%
    select(Pais, Gasto_medio_persona, Periodo) %>%
    filter(Periodo == "2016" & Pais != "Total")

gastos_medio_residencia_PN <- gastos_medio_residencia %>%
    select(Pais, Gasto_medio_persona, Periodo) %>%
    filter(Pais == "Países Nórdicos")

# Introducimos 4 filas iguales sobre los países nórdicos
# para separar con el join estos y poder representarlos en
# el mapa

gastos_medio_residencia <- rbind(gastos_medio_residencia, gastos_medio_residencia_
    gastos_medio_residencia_PN, gastos_medio_residencia_PN, gastos_medio_residenci

año_ref <- 2016

# Datos
countries <- gisco_get_countries(year = año_ref, resolution = 20) %>%
    select(CNTR_ID, NAME_ENGL, geometry) %>%
```

```
    st_transform(3035)

paises_english <- c("United Kingdom", "Denmark", "Germany", "France",
    "Italy", "Sweden", "Norway", "Iceland", "Finland")
countries_filtered <- countries %>%
    filter(NAME_ENGL %in% paises_english)

# Incluir columna de identificación de los países para unir
# con las coordenadas de los mapas

CNTR_ID <- c("UK", "DK", "DE", "FR", "IT", "SE", "NO", "IS",
    "FI")
gastos_medio_residencia <- cbind(gastos_medio_residencia, CNTR_ID)


gastos_medio_viz <- gastos_medio_residencia %>%
    left_join(countries, by = join_by(CNTR_ID == CNTR_ID))



# Mapa base
ggplot(gastos_medio_viz) + geom_sf(aes(geometry = geometry, fill = Gasto_medio_per
    xlim(c(2200000, 7150000)) + ylim(c(1380000, 5500000))
```



85

```
# Marta: Te comenté esta parte porque me estaba dando
# problemas al hacer knit no se por qué
```

```
# Gráfico ggplot(gastos_medio_viz) + # Primera capa con
# todos los países geom_sf(aes(geometry= geometry), data =
# gastos_medio_viz, fill = 'grey80', color = NA) + #
# Establece límites xlim(c(2200000, 7150000)) +
# ylim(c(1380000, 5500000))
```