*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: Máster en Ciencia de Datos, Universitá de Valencia.

**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.

```
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
library(ggiraph)  # Librería para añadir interactividad al mapa geográfico
library(GGally)  # Librería para ggpairs
```

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, Países Nórdicos, Reino Unido}:

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

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, Países Nórdicos, Reino Unido}:

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

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.

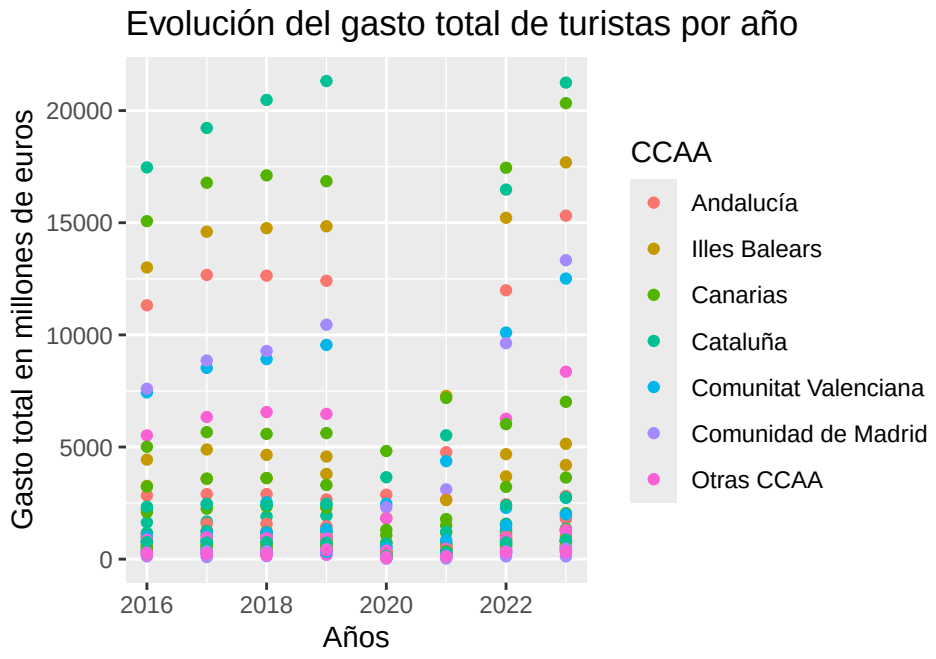
**Table 3.** Descripción de las variables

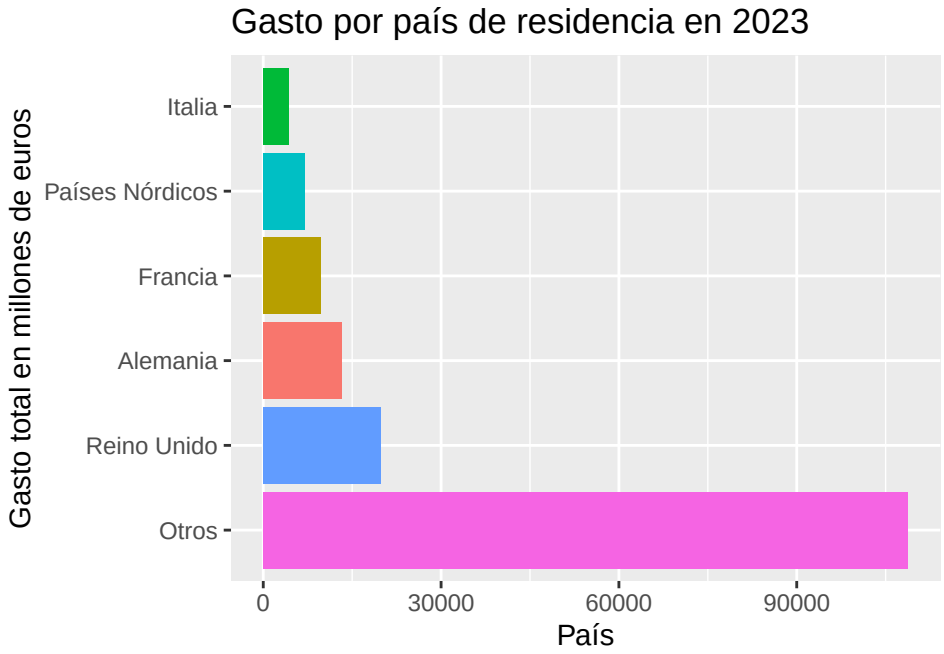
| Nombre variable | Unidad | Valores |
|---|---|---|
| Pais | - | Alemania, Italia, Países Nórdicos, Francia, Reino Unido, Otros |
| CCAA | - | Andalucía, Illes Balears, Canarias, Cataluña, Comunitat Valenciana, Comunidad de Madrid, Otras CCAA |
| Periodo | Año | 2016-2023 |
| Gasto_total | Millones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

Finalmente, antes de comenzar a visualizar datos y detectar patrones vemos un pequeño resumen de los datos:

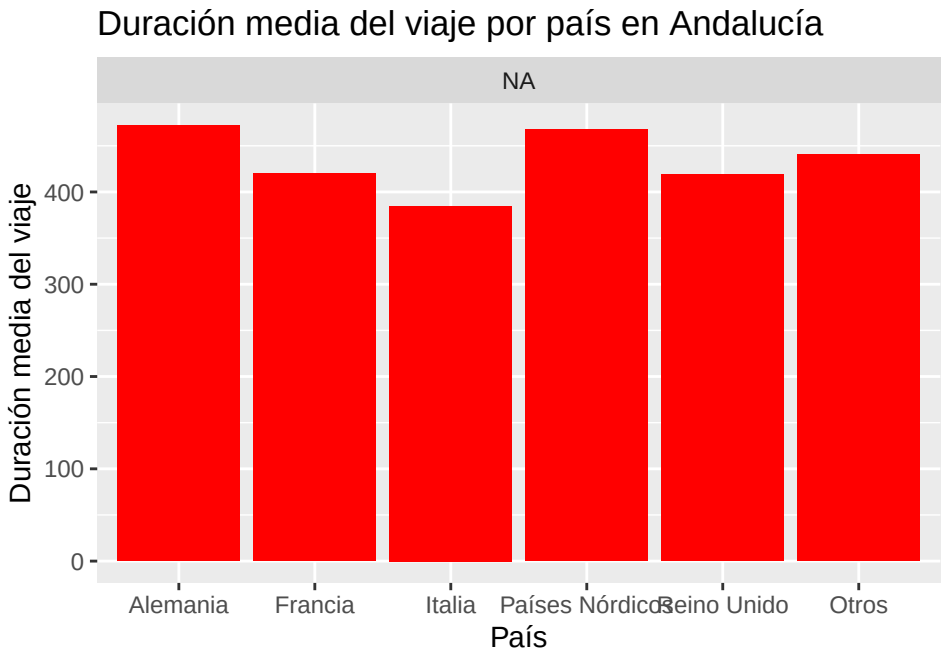
| Gasto total | Gasto medio por persona | Gasto medio diario por persona | Duración media |
|---|---|---|---|
| Min. : 22.28 | Min. : 383.0 | Min. : 56.0 | Min. : 3.610 |
| 1st Qu.: 396.06 | 1st Qu.: 838.2 | 1st Qu.:107.8 | 1st Qu.: 6.060 |
| Median : 857.80 | Median :1003.0 | Median :132.0 | Median : 7.630 |
| Mean : 2714.09 | Mean :1005.2 | Mean :135.9 | Mean : 7.756 |
| 3rd Qu.: 2562.09 | 3rd Qu.:1165.5 | 3rd Qu.:156.0 | 3rd Qu.: 9.198 |
| Max. :21318.75 | Max. :1743.9 | Max. :300.8 | Max. :14.270 |

## 5. Prueba de ggplot

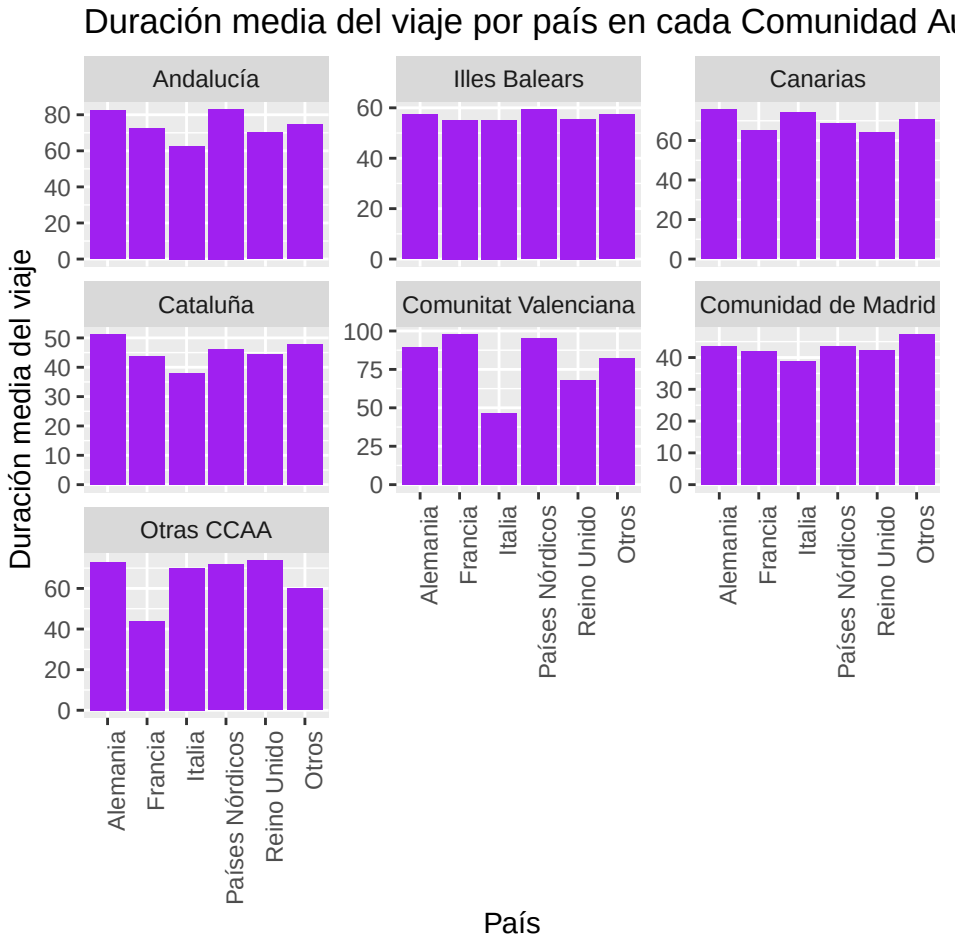



Gasto por país de residencia en 2023
Gasto total en millones de euros
Italia
Países Nórdicos
Francia
Alemania
Reino Unido
Otros
0
30000
60000
90000
País

75

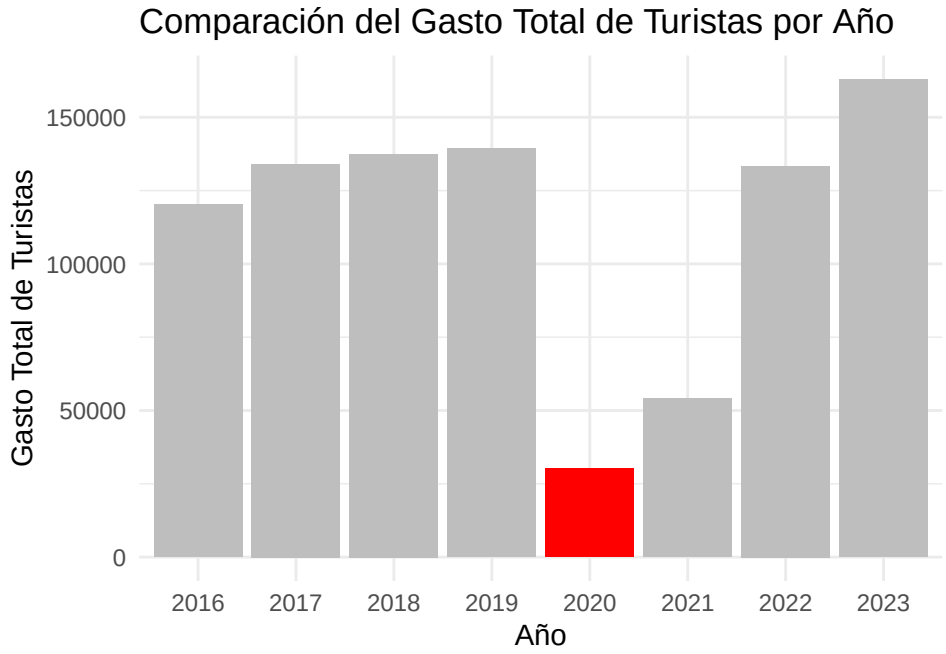

Duración media del viaje por país en Andalucía
NA
Duración media del viaje
400
300
200
100
0
Alemania
Francia
Italia
Países Nórdicos
Reino Unido
Otros
País

76

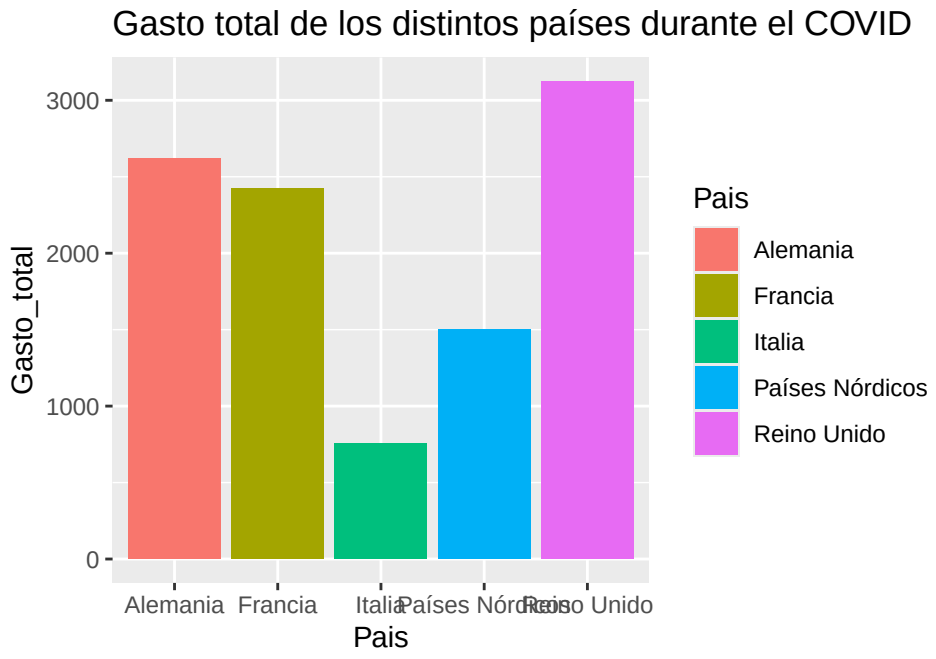



**6. Probamos a analizar el efecto del COVID en el gasto de los turistas, así como los países que más redujeron su gasto debido a la pandemia.**

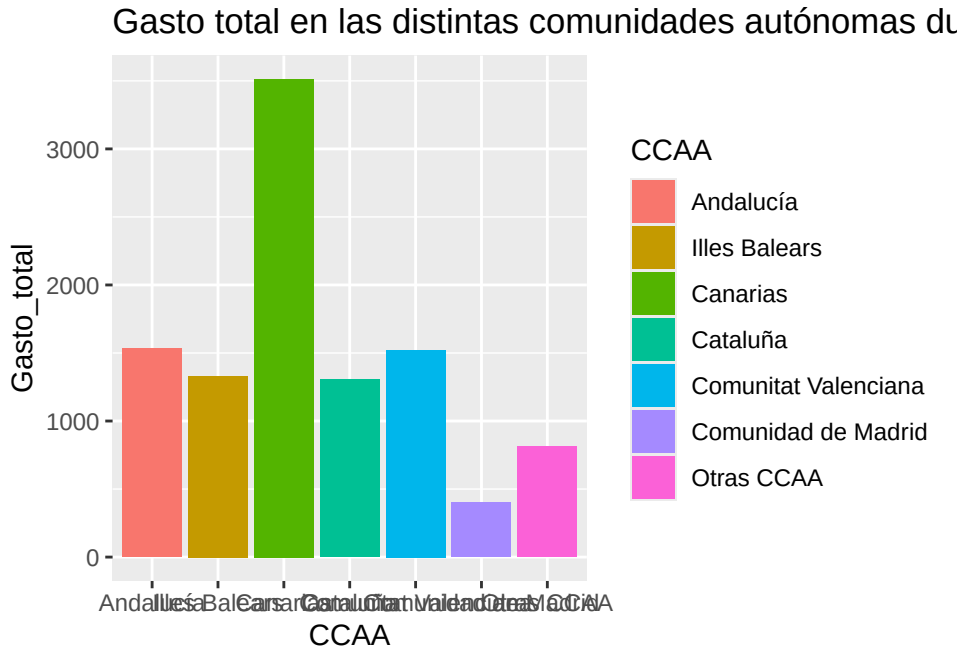
```
## La media del gasto total durante el COVID es de 719.3212 millones de euros

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


Comparación del Gasto Total de Turistas por Año
Gasto Total de Turistas
150000
100000
50000
0
2016
2017
2018
2019
2020
2021
2022
2023
Año

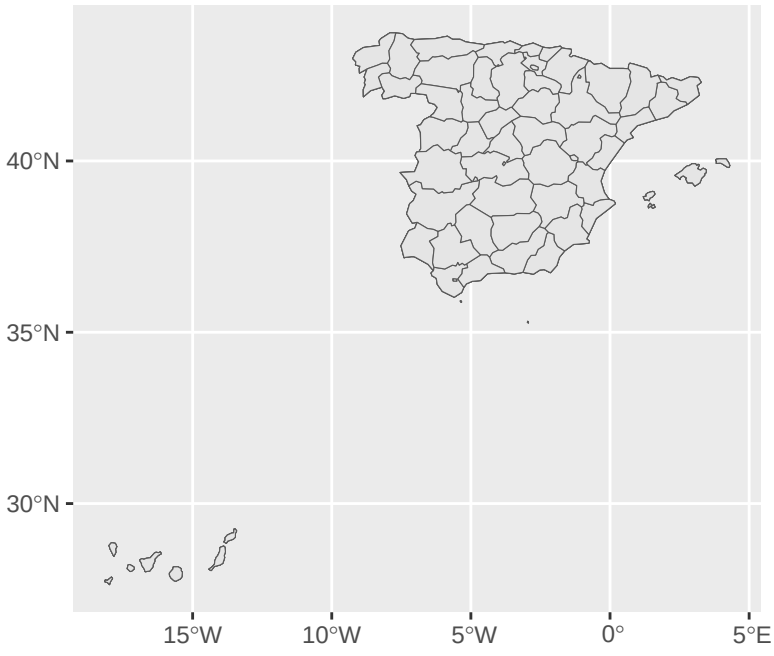
82


Gasto total de los distintos países durante el COVID
Gasto_total
3000
2000
1000
0
Alemania
Francia
Italia
Países Nórdicos
Reino Unido
Pais
Pais
Alemania
Francia
Italia
Países Nórdicos
Reino Unido

83




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




## 7. Búsqueda de relaciones entre variables cuantitativas

Realizamos una visualización previa para detectar relaciones entre variables cuantitativas. En el gráfico 1, se observan 3 relaciones que podemos estudiar un poco más a fondo: el gasto medio por persona frente al gasto medio diario, el gasto medio por persona frente a la duración media y el gasto medio diario frente a la duración media.

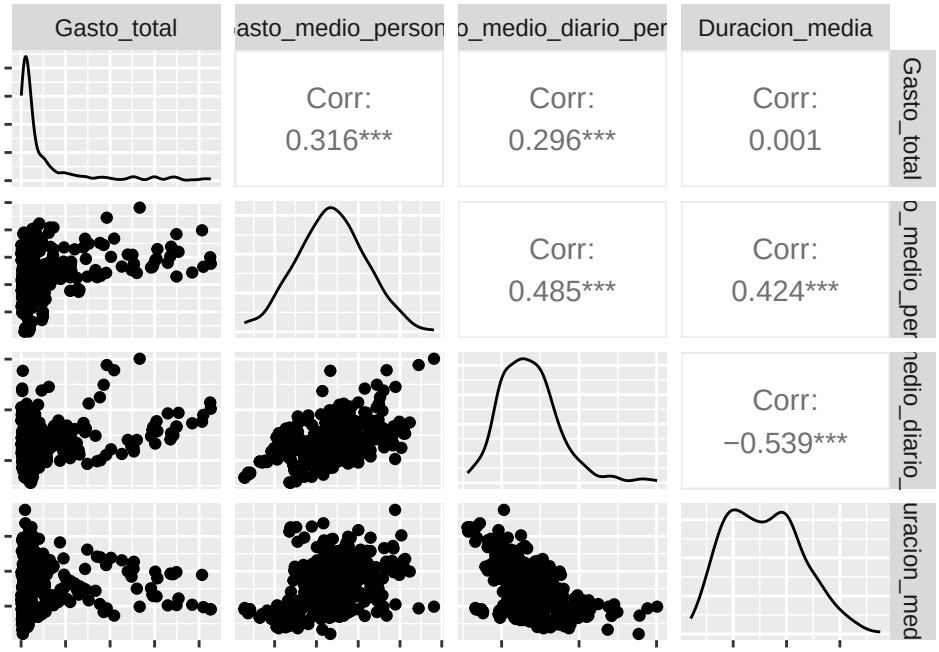



**Figure 1.** Gráficos de dispersión y correlaciones entre variables cuantitativas.

### *7.1. Gasto medio por persona y duración media*

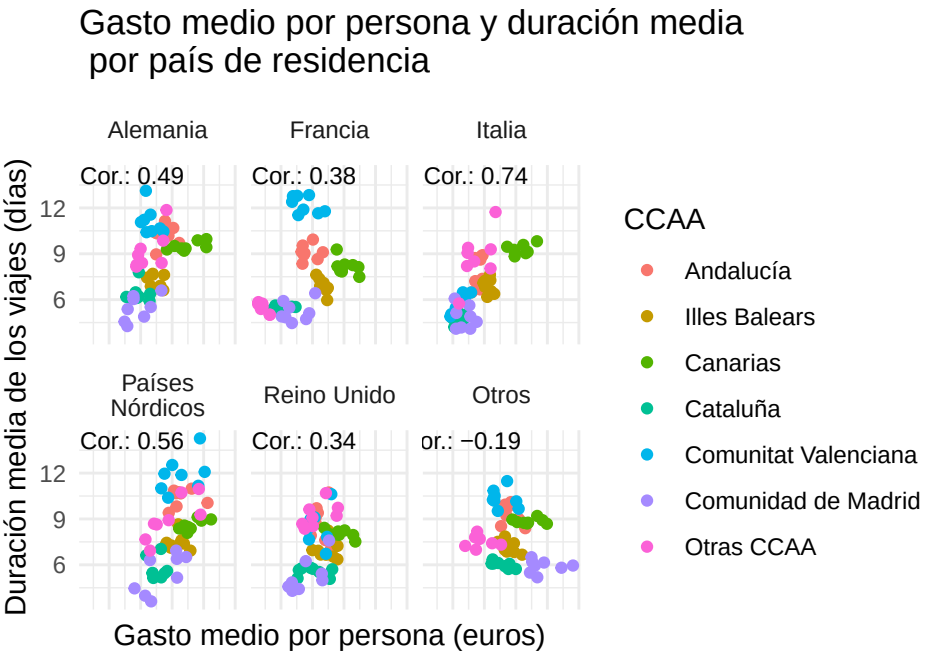



**Figure 2.** Gráficos de dispersión y correlaciones entre gasto medio por persona y la duración media de los viajes .

Se observa una relación aproximadamente lineal para la mayoría de países, donde se distinguen claramente la tendencia de duración de los viajes en las distintas comunidades autónomas, casi siempre mayor en la Comunitat Valenciana (salvo en Italia, probablemente por cercanía), excepto en la categoría de otros países. Debido a que esta categoría engloba a el resto de países del mundo, es más díficil encontrar relaciones.

### 7.2. *Gasto medio por persona frente al gasto medio diario por persona*

Esta parece a priori la relación más obvia entre variables.

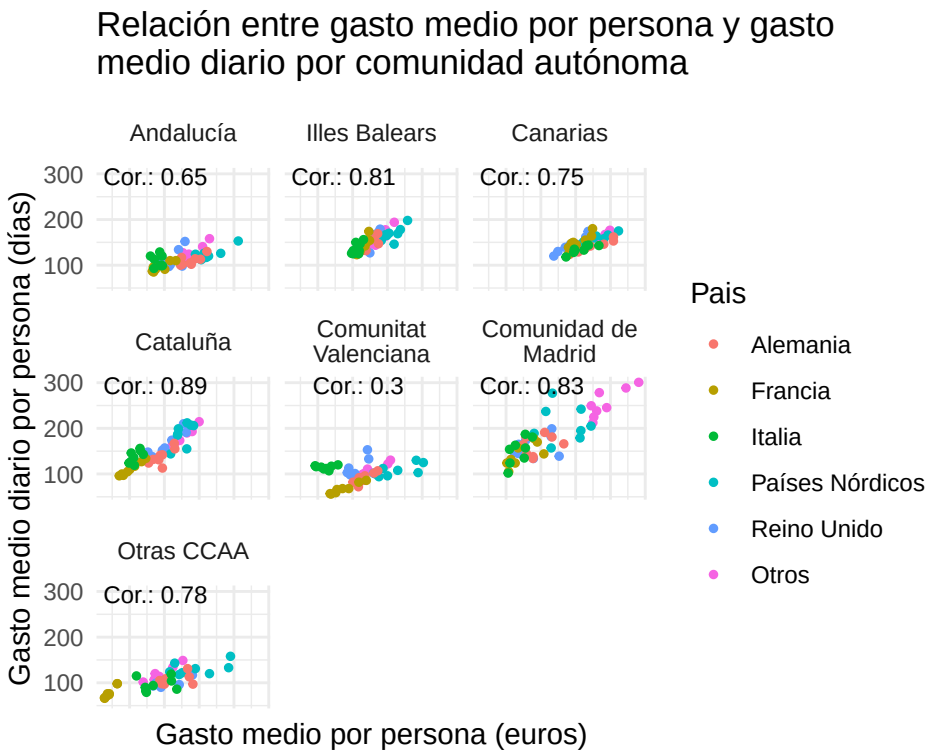



**Figure 3.** Gráficos de dispersión y correlaciones entre gasto medio por persona y gasto medio diario por persona .

En la Comunitat Valenciana es donde hay una peor correlación, esto puede estar relacionado con que es también el destino de mayor duración de los viajes, lo que puede provocar un menor gasto medio diario.

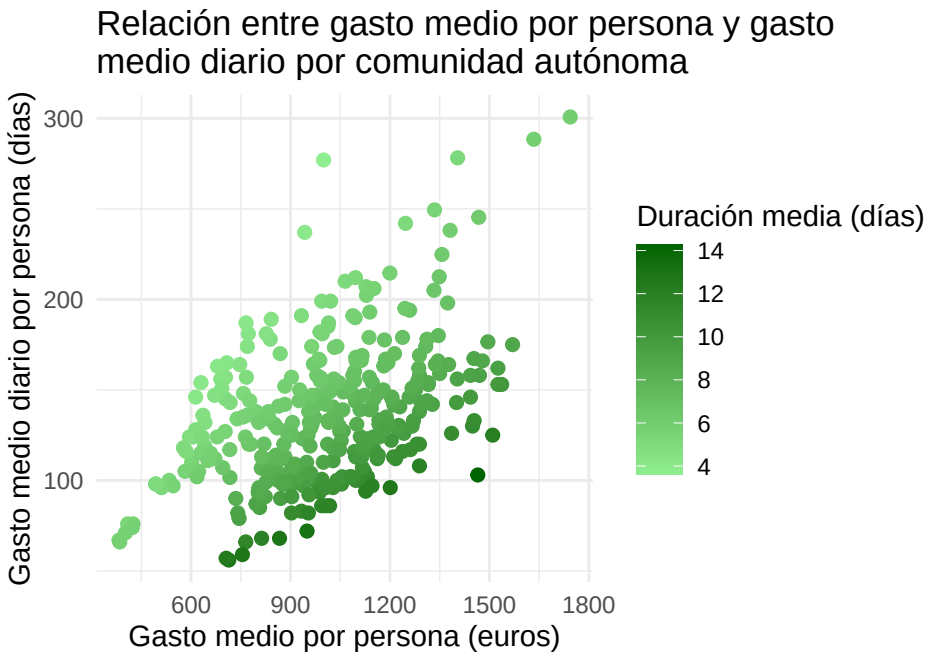



**Figure 4.** Gráfico de dispersión entre gasto medio por persona y gasto medio diario .

En el gráfico 4 se puede ver que generalmente a mayor gasto medio por persona también se produce un mayor gasto medio diario. Sin embargo, observamos que aparentemente un aumento de la duración de los viajes provoca una disminución del gasto medio diario. Vamos a visualizar esto más claramente a continuación.

### *7.3. Gasto medio diario y duracion media*

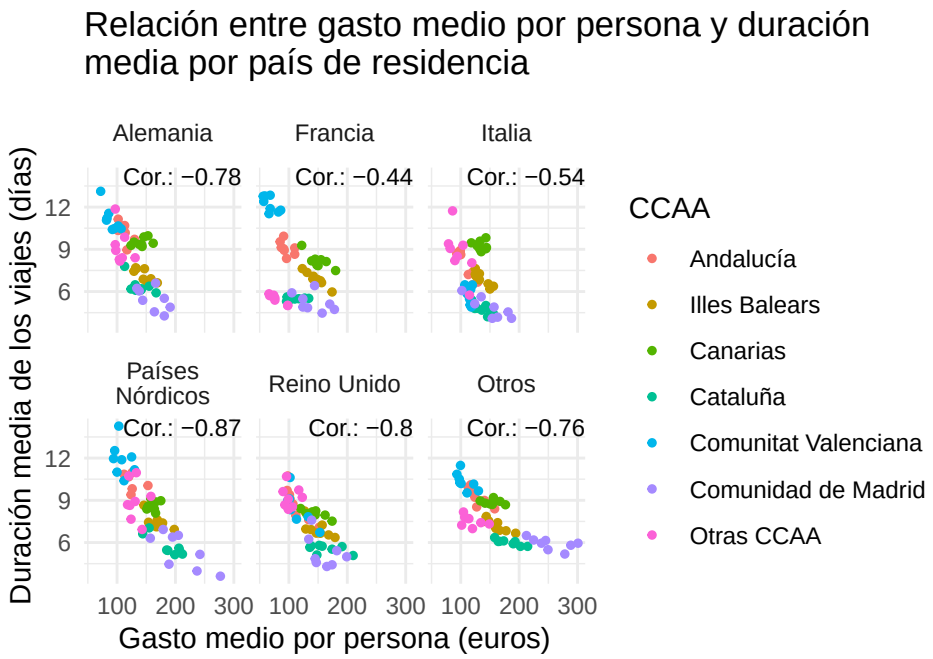



**Figure 5.** Gráfico de dispersión entre gasto medio diario por persona y duración media de los viajes .

Ahora podemos observar en el Gráfico 5 de manera clara que a mayor duración de los viajes se produce un menor gasto medio por persona.