MDPI

Article

Análisis de la relación entre la actividad turística y el empleo en el sector servicios en España

Iyán Álvarez, Azahara Martínez, Juan Alcaraz,

* Correspondence:

Abstract: El presente informe analiza la evolución del turismo en España y su relación con la ocupación en el sector servicios, a partir de datos oficiales del INE. Se integraron dos bases de datos —una sobre viajeros y pernoctaciones, y otra sobre empleo por sectores— para realizar un análisis exploratorio, univariante y bivariante a nivel de comunidad autónoma. Los resultados muestran una clara tendencia creciente y estacional en la actividad turística, así como una relación positiva y estadísticamente significativa entre el turismo y el empleo en la mayoría de las comunidades autónomas. Sin embargo, las elasticidades obtenidas son bajas, lo que indica que, aunque ambas variables evolucionan conjuntamente, el empleo responde de forma inelástica a las variaciones del turismo. En conjunto, los resultados sugieren que el turismo constituye un motor relevante del empleo en los servicios, aunque su efecto es moderado y depende de la estructura económica de cada territorio.

Citation: . Análisis de la relación entre la actividad turística y el empleo en el sector servicios en España. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

1. INTRODUCCIÓN

1.1. CONTEXTO GENERAL

El turismo constituye uno de los principales motores económicos de España, aportando una parte significativa al Producto Interior Bruto y al empleo nacional. La actividad turística no solo influye en la generación de riqueza, sino también en el dinamismo de otros sectores, como el transporte, la hostelería o el comercio.

En los últimos años, la evolución del turismo español ha experimentado variaciones notables, especialmente durante el periodo marcado por la pandemia de la COVID-19, que provocó una caída sin precedentes en los flujos de viajeros y en los niveles de ocupación del sector servicios. A partir de 2021, el proceso de recuperación ha sido desigual entre las comunidades autónomas, lo que hace necesario un análisis detallado que permita comprender las tendencias y relaciones entre las distintas variables implicadas.

En este contexto, el presente trabajo realiza un análisis exploratorio a partir de datos oficiales del **Instituto Nacional de Estadística (INE)**, con el objetivo de examinar la evolución del turismo y su relación con la ocupación en el sector servicios. Este enfoque permite no solo identificar patrones y correlaciones, sino también detectar posibles valores atípicos derivados de eventos excepcionales, como la crisis sanitaria, que pueden distorsionar la interpretación de los resultados.

1.2. DESCRIPCIÓN DE LOS DATOS

Para el desarrollo del presente análisis se han empleado dos bases de datos principales.

La primera recoge información sobre el turismo en España, desglosada por comunidades autónomas y provincias, e incluye variables como el número de viajeros y las pernoctaciones registradas a lo largo de varios años.

La segunda base de datos corresponde al sector servicios, concretamente a la ocupación laboral dentro del mismo, y se analiza con el propósito de explorar su relación con la evolución del turismo.

1.3. CUESTIONES A TRATAR

En este estudio se abordan **tres** cuestiones principales:

1. ¿Cómo tratar los valores atípicos (outliers) asociados al periodo de la pandemia de la COVID-19 (2019-2021)?
2. ¿Qué evolución ha experimentado el turismo en las distintas comunidades autónomas a lo largo del periodo analizado?
3. ¿Qué relación existe entre la evolución del turismo y la ocupación en el sector servicios?

2. IMPORTACIÓN

Para garantizar el uso de datos actualizados, la importación se realiza directamente desde las URLs oficiales del INE, evitando copias locales.

Antes de la carga definitiva, se comprobó la codificación de los archivos para evitar errores de lectura, optando por el formato **UTF-8**.

Finalmente, los datos se importaron y almacenaron en dos data frames principales: uno para turismo y otro para ocupación.

3. ANÁLISIS UNIVARIANTE

Antes de realizar el análisis univariante, es necesario limpiar y preparar las bases de datos. A continuación, se procede al tratamiento individual de cada una: primero la base de turismo y posteriormente la base de ocupación.

3.1. TURISMO

3.1.1. Limpieza

Una vez importados los datos, se analizan las características generales del conjunto correspondiente al turismo en España. Este dataset recoge el número de viajeros y pernoctaciones por comunidad y provincia a lo largo del tiempo.

En un inicio, este dataset se forma de 134.820 filas y 8 columnas, cuyo nombre resulta completamente representativo.

Las variables incluidas son:

- **Totales.Territoriales**: columna con un único valor (“Total Nacional”).
- **Comunidades.y.Ciudades.Autónomas**: variable categórica que representa las comunidades o ciudades autónomas de España.
- **Provincias**: variable categórica que muestra la provincia asociada.
- **Viajeros.y.pernoctaciones**: variable categórica. Indica el tipo de registro (viajeros o pernoctaciones).
- **Residencia..Nivel.1**: variable categórica redundante que contiene el valor “Total” en todas sus observaciones.
- **Residencia..Nivel.2**: variable categórica que indica el origen de los viajeros, diferenciando entre residentes en España y residentes en el extranjero.
- **Periodo**: variable tipo fecha que representa el año y mes de la observación.
- **Total**: valor numérico de viajeros o pernoctaciones expresado en millones.

Las variables Totales.Territoriales, Viajeros.y.pernoctaciones, Residencia..Nivel.1 y Residencia..Nivel.2, no aportan información relevante y se eliminan en fases posteriores.

A continuación, se utilizó la función `glimpse()` del paquete `dplyr` para obtener una vista general del dataset, comprobando que todas las columnas se importaron como tipo carácter y que será necesario ajustar sus tipos en etapas posteriores.

```
df_Turismo %>% glimpse()
```

```
## Rows: 134,820
## Columns: 8
## $ Totales.Territoriales            <chr> "Total Nacional", "Total Nacional", "~
## $ Comunidades.y.Ciudades.Autónomas <chr> "", "", "", "", "", "", "", "", "", "~
## $ Provincias                       <chr> "", "", "", "", "", "", "", "", "", "~
## $ Viajeros.y.pernoctaciones        <chr> "Viajero", "Viajero", "Viajero", "Via~
## $ Residencia..Nivel.1              <chr> "Total", "Total", "Total", "Total", "~
## $ Residencia..Nivel.2              <chr> "", "", "", "", "", "", "", "", "", "~
## $ Periodo                          <chr> "2025M09", "2025M08", "2025M07", "202~
## $ Total                            <chr> "12.050.972", "13.828.672", "13.044.7~
```

Gracias a esto, se observa que todas las columnas se importaron inicialmente como tipo character. Por ello, en la fase de preparación se realizará la conversión a tipos de datos adecuados (fechas, factores y numéricos) para facilitar el análisis estadístico y gráfico.

Para asegurar la correcta manipulación numérica, se eliminaron los puntos utilizados como separadores de miles en la columna Total.

```
df_Turismo <- df_Turismo %>%
  mutate(Total = str_replace_all(Total, "\\.", ""))
```

Posteriormente, se aplicó `sapply()` junto con `unique()` para revisar los valores de cada variable y detectar posibles inconsistencias.

```
sapply(df_Turismo, function(x) length(unique(x)))
```

```
##            Totales.Territoriales Comunidades.y.Ciudades.Autónomas
##                                1                                20
##                        Provincias        Viajeros.y.pernoctaciones
##                               51                                 2
##              Residencia..Nivel.1              Residencia..Nivel.2
##                                1                                 3
##                          Periodo                             Total
##                              321                             95210
```

Se identificaron varias **incidencias**: la columna Totales.Territoriales presentaba un único valor (“Total nacional”), y Comunidades.y.Ciudades.Autónomas incluía 20 categorías en lugar de 19, debido a registros vacíos o duplicados.

```
print(unique(df_Turismo$Comunidades.y.Ciudades.Autónomas))
```

```
##  [1] ""                          "01 Andalucía"
##  [3] "02 Aragón"                 "03 Asturias, Principado de"
##  [5] "04 Balears, Illes"         "05 Canarias"
##  [7] "06 Cantabria"              "07 Castilla y León"
##  [9] "08 Castilla - La Mancha"   "09 Cataluña"
## [11] "10 Comunitat Valenciana"   "11 Extremadura"
## [13] "12 Galicia"                "13 Madrid, Comunidad de"
## [15] "14 Murcia, Región de"      "15 Navarra, Comunidad Foral de"
## [17] "16 País Vasco"             "17 Rioja, La"
## [19] "18 Ceuta"                  "19 Melilla"
```

Además, se observó que la variable CCAA presentaba registros vacíos o en blanco, lo que explica las discrepancias detectadas anteriormente. Estos valores se eliminarán en la fase de limpieza del dataset.

El siguiente paso fue eliminar las filas y columnas redundantes, así como los valores en blanco, manteniendo únicamente la información agregada por comunidades autónomas.

```
df_Turismo <- df_Turismo %>%
  select(-Totales.Territoriales) %>%
  filter(str_trim(Comunidades.y.Ciudades.Autónomas) != "") %>%
  arrange(Comunidades.y.Ciudades.Autónomas, Provincias, Periodo)
```

Posteriormente, con el fin de facilitar la interpretación y el análisis comparativo, se optó por trabajar con los datos agregados a nivel de comunidades y ciudades autónomas. Este enfoque proporciona una visión más general del comportamiento del turismo en España, adecuada para su posterior comparación con la ocupación en el sector servicios.

En consecuencia, se eliminaron las variables no relevantes para este nivel de análisis (Provincias, Viajeros.y.pernoctaciones, Residencia..Nivel.1 y Residencia..Nivel.2).

```
df_Turismo <- df_Turismo %>%
  select(-c(Provincias,Viajeros.y.pernoctaciones,
            Residencia..Nivel.1,
            Residencia..Nivel.2))
```

Dado que todas las variables se importaron inicialmente como texto, fue necesario convertir Periodo al formato de fecha. Para ello, se sustituyó el carácter "M" por un guion y se aplicó la función `ym()` para reconocer correctamente el año y el mes.

```
df_Turismo <- df_Turismo %>%
  mutate(Periodo = str_replace(Periodo, "M", "-"),
    Periodo = ym(Periodo))
```

Además, la variable Total se convirtió a tipo numérico para permitir la realización de cálculos y análisis estadísticos sin errores de formato.

```
df_Turismo <- df_Turismo %>%
  mutate(Total=as.numeric(Total))
```

Antes de convertir la variable Comunidades.y.Ciudades.Autónomas a tipo factor, se limpiaron y estandarizaron sus valores eliminando números, espacios y caracteres innecesarios.

Finalmente, se renombró como CCAA y se definió como variable categórica tipo factor.

```
df_Turismo <- df_Turismo %>%
  rename(CCAA = Comunidades.y.Ciudades.Autónomas) %>%
  mutate(
    CCAA = str_replace(CCAA, "^[0-9]+\\s+", ""),
    CCAA = str_replace(CCAA,
                       "^\\s*([^,]+?)\\s*,\\s*([^,]+?)\\s*$", "\\2 \\1"),
    CCAA = str_squish(CCAA)
  ) %>%
  mutate(CCAA = as.factor(CCAA))
```

Para continuar con el análisis, se creó un nuevo conjunto de datos agregado a nivel de comunidad autónoma.

```
df_Turismo <- df_Turismo %>%
  group_by(CCAA, Periodo) %>%
  summarise(Total = sum(Total, na.rm = TRUE), .groups = "drop")
```

Tras la limpieza, se comprobó la ausencia de espacios en blanco y valores faltantes (NA) para asegurar la integridad del conjunto de datos. Al no detectarse incidencias, el dataset quedó listo para el análisis y se guardó en formato RData.

3.1.2. Detección de outliers

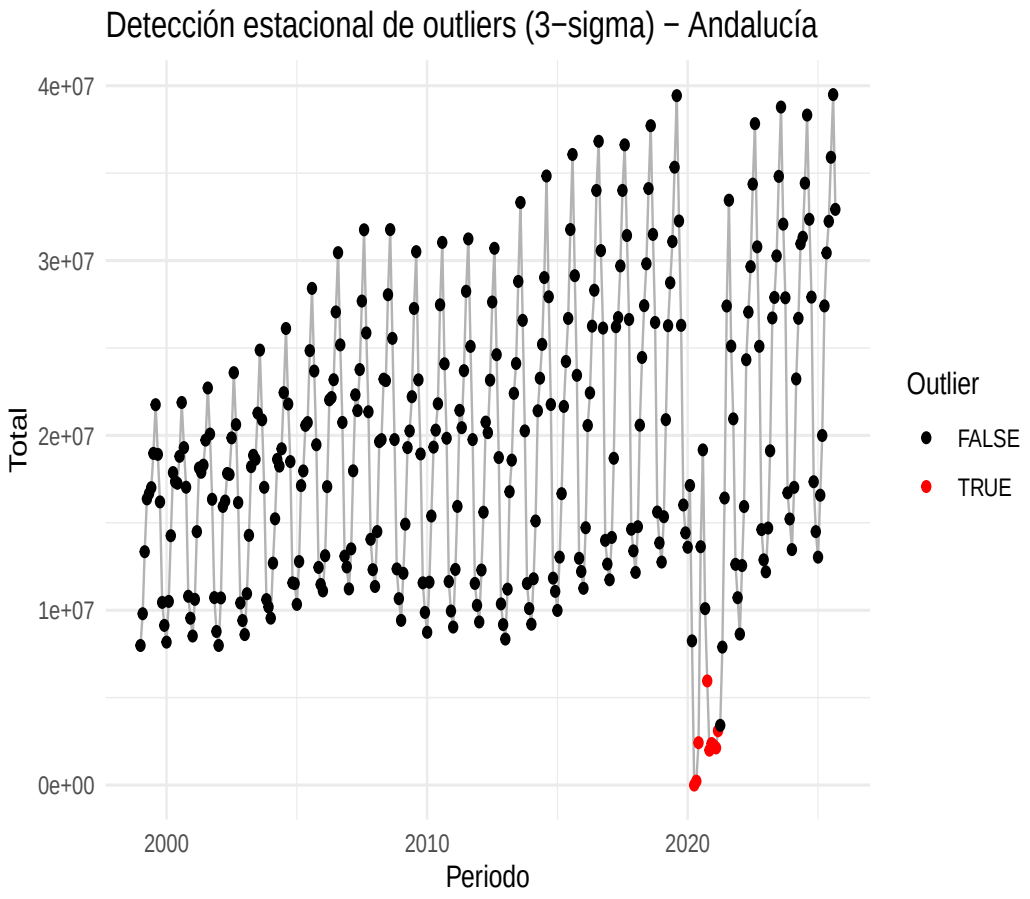
Para la detección de valores atípicos en las series de turismo se aplicaron tres métodos estadísticos complementarios:

- Regla 3-sigmas: considera como *outliers* los valores que se alejan más de tres desviaciones típicas respecto a la media.
- Método de Hampel: utiliza la mediana y la desviación absoluta mediana (MAD) para detectar observaciones atípicas, siendo más robusto frente a valores extremos.
- Método del Boxplot: identifica *outliers* a partir de los límites del rango intercuartílico (1,5×IQR).

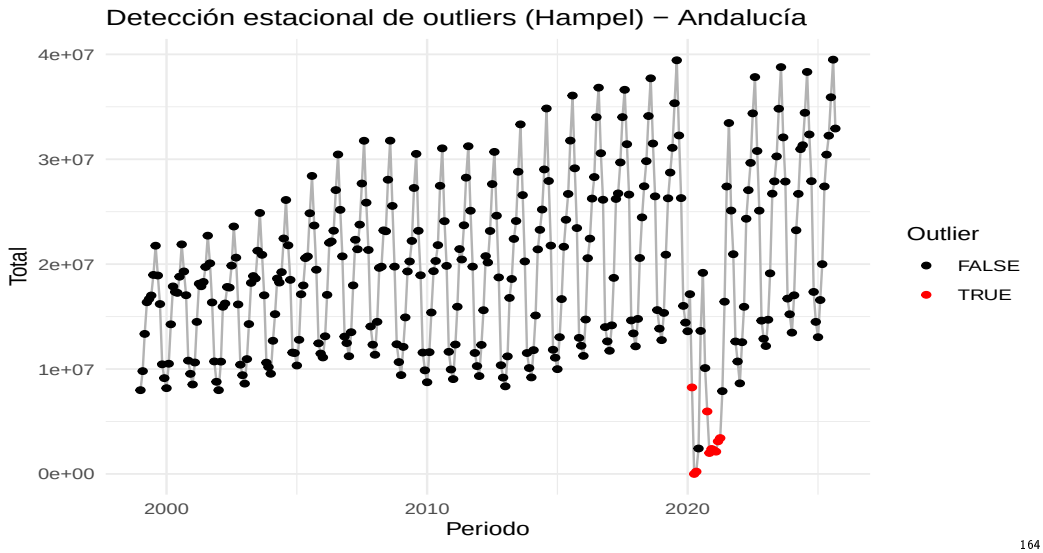
En todos los casos se aplicó una versión estacional del método, calculando los parámetros (media, desviación, mediana o cuartiles) por mes y comunidad autónoma, con el objetivo de evitar clasificar como atípicos los picos o caídas debidos a la estacionalidad turística.

Los gráficos se generaron automáticamente para cada comunidad autónoma, aunque solo se muestra un ejemplo en el informe. Todos ellos se han guardado en carpetas independientes:
`graficos_outliers_3sigma`, `graficos_outliers_hampel` y `graficos_outliers_boxplot`.

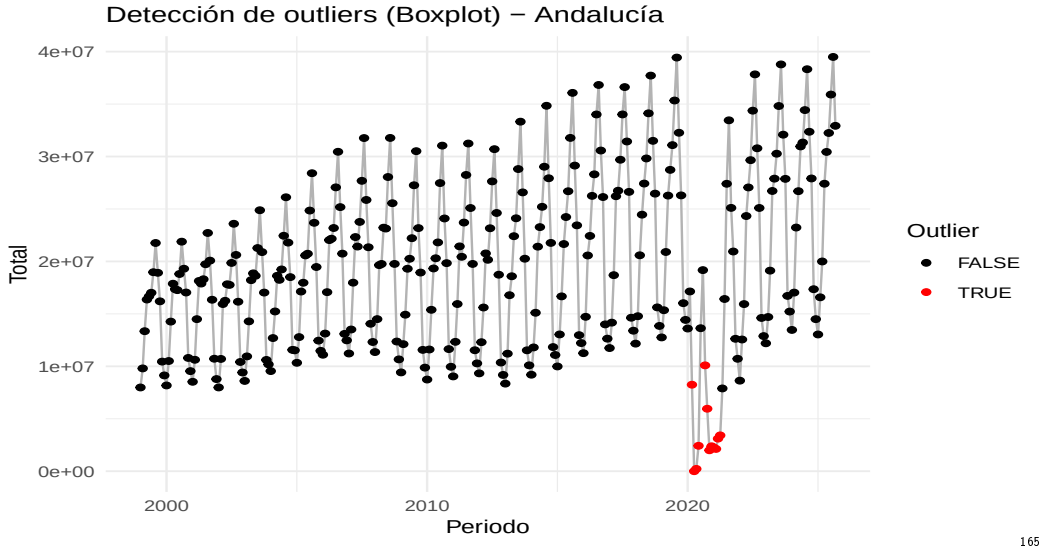
```
Sigma3_est <- function(x, k = 3) {
  media <- mean(x, na.rm = TRUE)
  desv <- sd(x, na.rm = TRUE)
  out <- abs(x - media) > (k * desv)
}
```



```r
Hampel_est <- function(x, k = 3) {
  abs(x - median(x)) > k * mad(x)
}
```




```r
Boxplot_est <- function(x, k = 1.5) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  inf <- Q1 - k * IQR_value
  sup <- Q3 + k * IQR_value
  (x < inf) | (x > sup)
}
```




```
## # A tibble: 19 x 4
##    CCAA                      n_hampel n_boxplot n_sigma3
##    <fct>                        <int>     <int>    <int>
##  1 Ceuta                           20        21        8
##  2 Illes Balears                   18        17        8
##  3 Castilla - La Mancha            14        16       11
##  4 Castilla y León                 13        14        9
```

```
##  5 Comunitat Valenciana            12        13        10
##  6 La Rioja                        12        15        10
##  7 Región de Murcia                12        13         8
##  8 Cantabria                       11        13         7
##  9 Andalucía                       10        12         9
## 10 Extremadura                     10        11         7
## 11 Aragón                           9         9         6
## 12 Galicia                          9        12         6
## 13 Melilla                          7         4         1
## 14 Principado de Asturias           6         9         3
## 15 Cataluña                         5         6         3
## 16 Comunidad Foral de Navarra       1         1         0
## 17 Canarias                         0         0         0
## 18 Comunidad de Madrid              0         0         0
## 19 País Vasco                       0         0         0
```

Durante la aplicación de los distintos métodos de detección de valores atípicos (3-sigma, Hampel y Boxplot) se observó que algunas comunidades autónomas no presentaban *outliers*, lo cual resultaba poco coherente con la evolución visible de sus series temporales.

Para contrastar esta situación, se analizaron los estadísticos descriptivos de los valores mensuales (media, varianza y desviación típica). Este análisis permitió comprobar que, en algunos casos, como el de la Comunidad de Madrid, la desviación típica estacional es elevada debido a un crecimiento sostenido en los últimos años.

Esto implica que, aunque la variabilidad dentro de cada año es baja, la variabilidad entre estaciones es alta, provocando que el método 3-sigma no identifique correctamente ciertos valores que visualmente podrían considerarse atípicos.

Este comportamiento también se repite en otras comunidades de similares características, donde las medias y la dispersión entre estaciones son altas.

Dado que los resultados sugieren que los métodos tradicionales no captan adecuadamente todas las anomalías, y teniendo en cuenta el conocimiento contextual del fenómeno turístico, se ha decidido considerar como valores atípicos comunes el periodo comprendido entre marzo de 2020 y febrero de 2021, correspondiente al impacto más severo de la pandemia de la COVID-19.

Este enfoque permite homogeneizar el tratamiento de outliers en todas las comunidades, evitando que diferencias en la dispersión estacional alteren la comparación entre regiones.

3.1.3. Análisis de series de tiempo

Los datos de turismo referentes al periodo de la pandemia son claramente *outlier*, la pregunta principal a la que se debe hacer frente es si eliminarlo o si tenerlos en cuenta para posteriores análisis. Esta pregunta no es sencilla pues para ciertas CCAA, como se ha mencionado, hay métodos de detección de *outliers* que no los clasifican como tales.

La idea para justificar qué debemos hacer con estos datos anómalos es la siguiente, en general, si representamos la serie temporal de una comunidad autónoma se puede ver que la gran bajada del número de pernoctaciones a partir del mes de marzo de 2020, lo cual está directamente relacionado con el inicio del estado de alarma y la cuarentena nacional.

Tras esto se mantienen unas cifras muy bajas hasta aproximadamente dos años después (dependiendo de la comunidad), coincidiendo con la declaración del final de la pandemia por parte de la OMS, los datos de estancias vuelven prácticamente a los niveles prepandemia, lo cual sugiere que aunque la pandemia afectó de forma muy negativa al sector de la hostelería y el turismo, también este sector recuperó de forma bastante rápida cuando la situación sanitaria lo permitió.

Esto hace pensar que la pandemia simplemente puso una “pausa” al turismo y que tras esta pausa la actividad en este sector continuó desarrollándose de la misma forma que lo venía haciendo en los últimos años, aunque si es cierto que el crecimiento se desaceleró

ligeramente. Lo cual sugiere que los datos de la pandemia no aportan ningún tipo de información relevante que pueda ayudar a predecir o explicar el futuro del sector del turismo.

Pues no solo no aportan información relevante, si no que además añadirían ruido a los datos afectando de forma negativa a los posibles análisis que puedan hacerse.

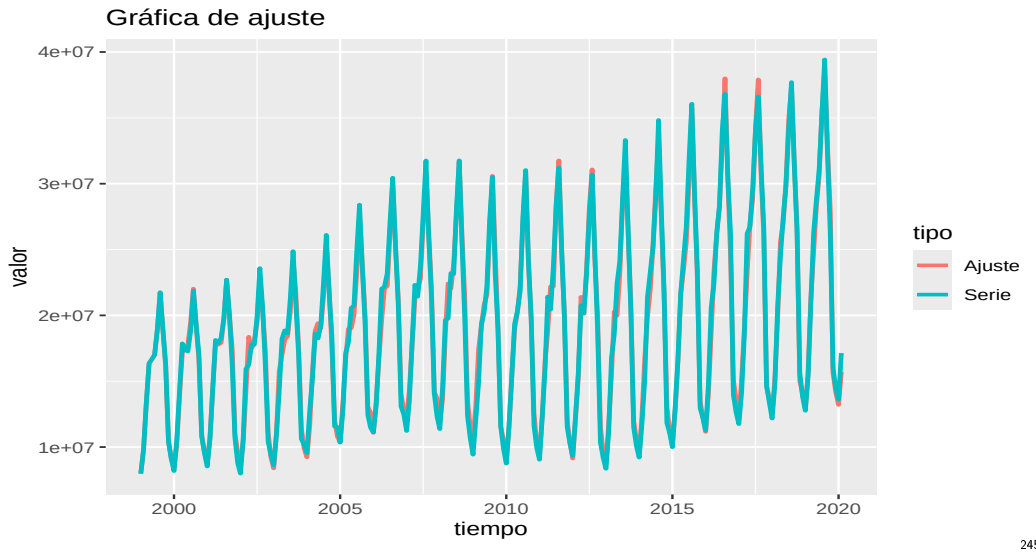
Para darle respaldo a esta hipótesis nos proponemos hacer un análisis de la serie de datos de cada CCAA, cosa que es muy natural por la forma en la que están organizados los datos.

Entonces, se tomarán los datos previos a la pandemia y se utilizarán para crear un modelo $ARIMA(p,d,q)(P,D,Q)$, el ajuste de los parámetros se hace de forma automática con las función `auto.arima`. Tras el ajuste se usa la función `forecast` que permite utilizar el modelo para predecir una serie de valores en el futuro, en este caso de predicen los 12 meses siguientes.

Los resultados para cada comunidad se guardaron en un archivo llamado `AnalisisBox.rsd`. Se trata de una lista, si queremos acceder a ellos se pueden importar con el comando `readRSD("AnálisisBox.rsd")` y asignándolo a una variable.

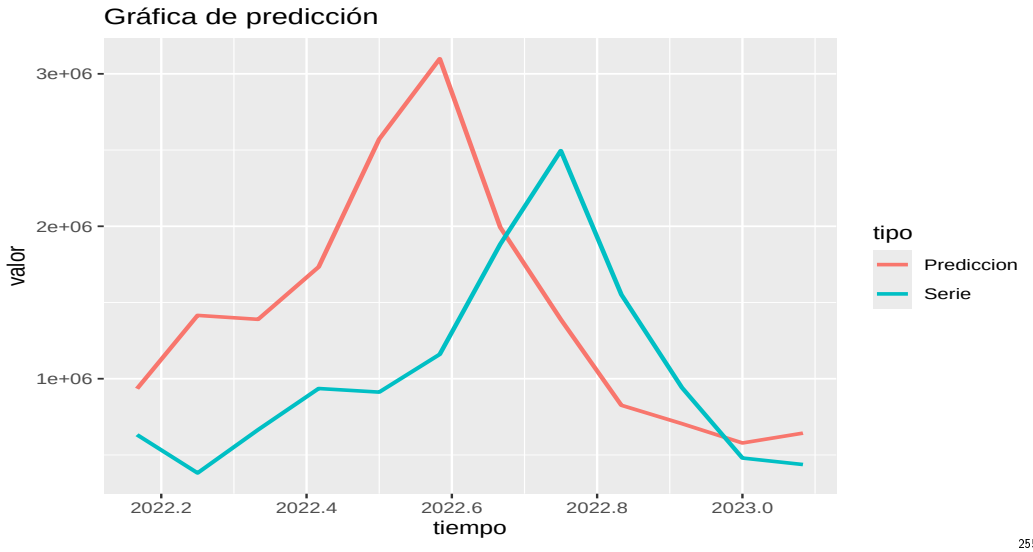
Se pueden ver las dierentes gráficas de juste para cada comunidad, como son muchos gráficos se muestra solo uno, pero se puede ver que el resto muestran resultados similares. Ponemos el caso de **Andalucía**

```
AnalisisBox[[2]]
```

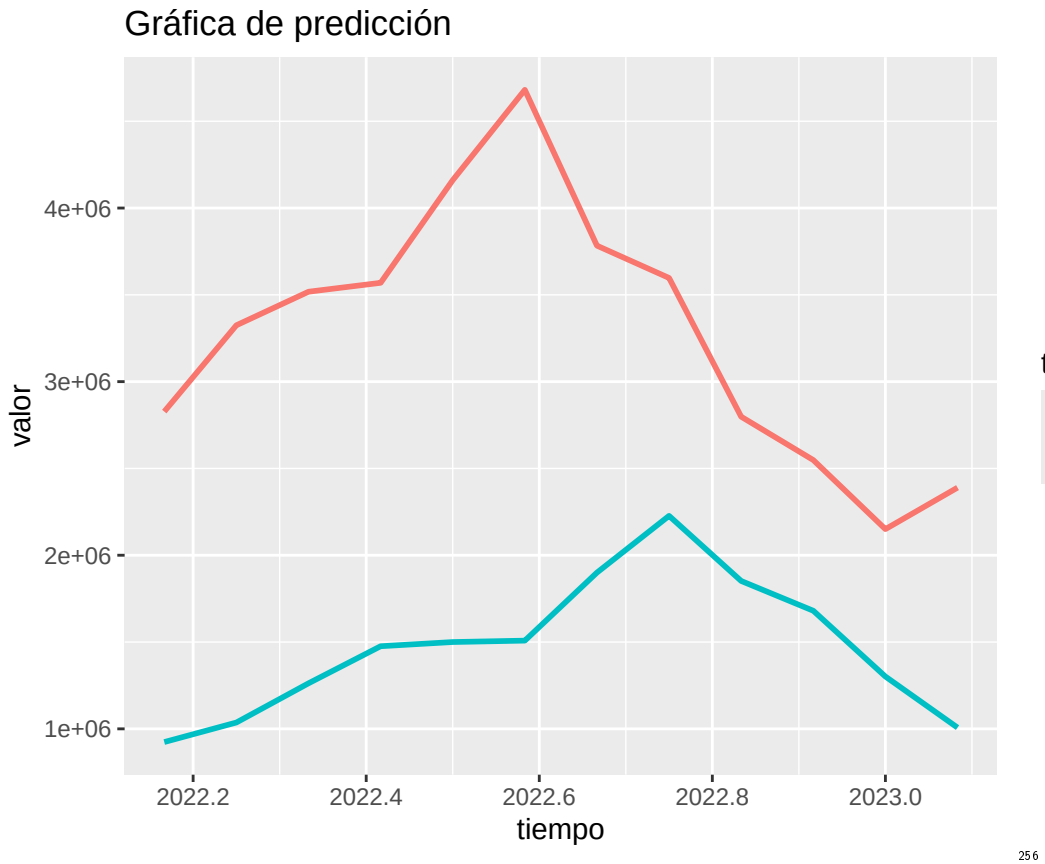


Entonces, tras ver que el modelo aproxima de forma razonable los datos prepandemia, hay razones para confiar en las predicciones, las cuales se pueden interpretar como los datos de pernoctaciones que hubieran sucedido en esos meses siguientes a febrero de 2020 si la pandemia no hubiera azotado.

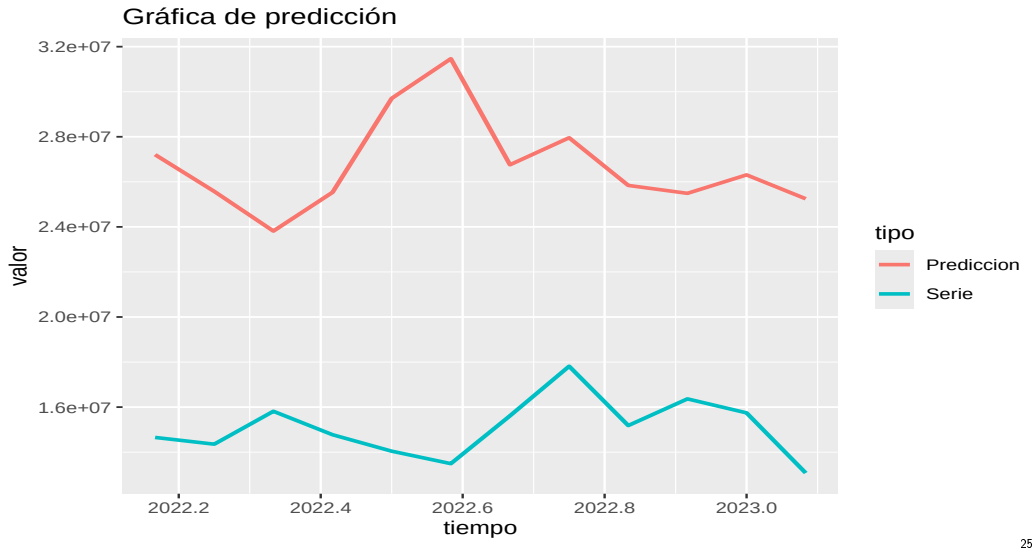
Lo más interesante ocurre si comparamos esas predicciones con los datos posteriores a la pandemia. Si bien es cierto que hay algunas comunidades para las cuales las predicciones no se ajustan en nada (son los casos de las ciudades autónomas) en el resto de casos se puede ver como la predicción y el valor postpandemia son iguales solo que con cifras algo menores menores.


Gráfica de predicción
3e+06
2e+06
1e+06
valor
2022.2
2022.4
2022.6
2022.8
2023.0
tiempo
tipo
Prediccion
Serie

255


Gráfica de predicción
4e+06
3e+06
2e+06
1e+06
valor
2022.2
2022.4
2022.6
2022.8
2023.0
tiempo
tipo

256



3.1.4. Análisis

Como primer paso del análisis univariante, se calcularon los estadísticos descriptivos básicos con el fin de obtener una visión general de las variables del conjunto de datos.

```
summary(df_Turismo)
```

```
##                     CCAA          Periodo                Total
##  Andalucía            : 321   Min.   :1999-01-01   Min.   :       0
##  Aragón               : 321   1st Qu.:2005-09-01   1st Qu.:  818367
##  Canarias             : 321   Median :2012-05-01   Median : 1978968
##  Cantabria            : 321   Mean   :2012-05-01   Mean   : 6265921
##  Castilla - La Mancha: 321   3rd Qu.:2019-01-01   3rd Qu.: 7983666
##  Castilla y León      : 321   Max.   :2025-09-01   Max.   :54032304
##  (Other)              :4173
```

El resumen estadístico muestra que todas las comunidades autónomas tienen el mismo número de observaciones, con registros trimestrales entre 1999 y 2025. La variable Total presenta un mínimo de 0 (probablemente asociado a la pandemia), una media de 6,27 millones y un máximo de 54 millones, reflejando una notable variabilidad entre regiones y periodos.

Dado que la función `summary()` ofrece información limitada para variables categóricas, se amplió el análisis calculando los principales estadísticos descriptivos de Total por comunidad autónoma. Este enfoque permite comparar la magnitud y variabilidad del turismo entre regiones, antes de profundizar en las relaciones en el análisis bivariante.

```
df_Turismo %>%
  group_by(CCAA) %>%
  summarise(
    media = mean(Total, na.rm = TRUE),
    mediana = median(Total, na.rm = TRUE),
    sd = sd(Total, na.rm = TRUE),
    min = min(Total, na.rm = TRUE),
    max = max(Total, na.rm = TRUE),
    IQR=IQR(Total, na.rm=TRUE),
    N = n()
  ) %>%
  arrange(desc(media)) %>%
  print()
```

```
## # A tibble: 19 x 8
##    CCAA                          media mediana     sd   min     max     IQR
##    <fct>                         <dbl>   <dbl>  <dbl> <dbl>   <dbl>   <dbl>
##  1 Cataluña                   20162927.  1.82e7 1.11e7     0 4.62e7 1.74e7
##  2 Illes Balears              19915553.  1.41e7 1.73e7     0 5.40e7 3.30e7
##  3 Canarias                   19889672.  2.02e7 6.40e6     0 3.25e7 9.84e6
##  4 Andalucía                  19415478.  1.88e7 8.28e6     0 3.95e7 1.29e7
##  5 Comunitat Valenciana       10438034.  1.00e7 4.12e6     0 2.10e7 5.60e6
##  6 Comunidad de Madrid         9063390.  9.19e6 2.98e6     0 1.51e7 4.74e6
##  7 Castilla y León             3752446.  3.81e6 1.26e6     0 6.91e6 1.75e6
##  8 Galicia                     3739708.  3.21e6 1.98e6     0 9.75e6 2.63e6
##  9 País Vasco                  2295078.  2.11e6 1.00e6     0 5.41e6 1.33e6
## 10 Aragón                      2254042.  2.18e6 7.28e5     0 4.60e6 8.69e5
## 11 Castilla - La Mancha        1767775.  1.82e6 4.56e5     0 2.81e6 6.30e5
## 12 Principado de Asturias      1506391.  1.33e6 8.70e5     0 4.02e6 1.11e6
## 13 Región de Murcia            1270981.  1.27e6 4.45e5     0 2.54e6 6.22e5
## 14 Cantabria                   1212464.  1.04e6 7.35e5     0 3.19e6 9.70e5
## 15 Extremadura                 1078914.  1.09e6 3.26e5     0 1.89e6 4.72e5
## 16 Comunidad Foral de Navarra   766060.  7.27e5 3.13e5     0 1.63e6 4.71e5
## 17 La Rioja                     458772.  4.65e5 1.54e5     0 9.40e5 2.12e5
## 18 Ceuta                         37138.  3.78e4 9.27e3     0 6.47e4 9.14e3
## 19 Melilla                       27683.  2.80e4 8.34e3     0 4.88e4 1.12e4
##    N
##  <int>
##  1 321
##  2 321
##  3 321
##  4 321
##  5 321
##  6 321
##  7 321
##  8 321
##  9 321
## 10 321
## 11 321
## 12 321
## 13 321
## 14 321
## 15 321
## 16 321
## 17 321
## 18 321
## 19 321
```

Los resultados muestran que Cataluña presenta la media más alta en el número de turistas, lo que refleja su consolidación como uno de los principales destinos turísticos del país. Le siguen Baleares, Canarias y Andalucía, todas ellas comunidades con una fuerte especialización en el sector turístico.

En el extremo opuesto, Melilla, Ceuta, La Rioja y Navarra registran las medias más bajas, lo que podría deberse a su menor capacidad de atracción turística o a un peso económico menos centrado en este sector.

Para obtener una visión general del comportamiento del turismo en el conjunto del país, se calcularon los principales estadísticos descriptivos del Total Nacional, agregando los valores de todas las comunidades por periodo.

```
nacional_mensual <- df_Turismo %>%
  group_by(Periodo) %>%
  summarise(Total = sum(Total, na.rm = TRUE), .groups = "drop")

nacional_mensual %>%
  summarise(
    media = mean(Total, na.rm = TRUE),
    mediana = median(Total, na.rm = TRUE),
    sd = sd(Total, na.rm = TRUE),
    var = var(Total, na.rm = TRUE),
    min = min(Total, na.rm = TRUE),
    max = max(Total, na.rm = TRUE),
    N = n()
  )
```

```
## # A tibble: 1 x 7
##        media   mediana        sd      var   min       max     N
##        <dbl>     <dbl>     <dbl>    <dbl> <dbl>     <dbl> <int>
## 1 119052505. 111782620 52682662. 2.78e15     0 247949935   321
```

Los resultados agregados a nivel nacional muestran una media de aproximadamente 119 millones de turistas por periodo y una mediana cercana a 112 millones, lo que indica

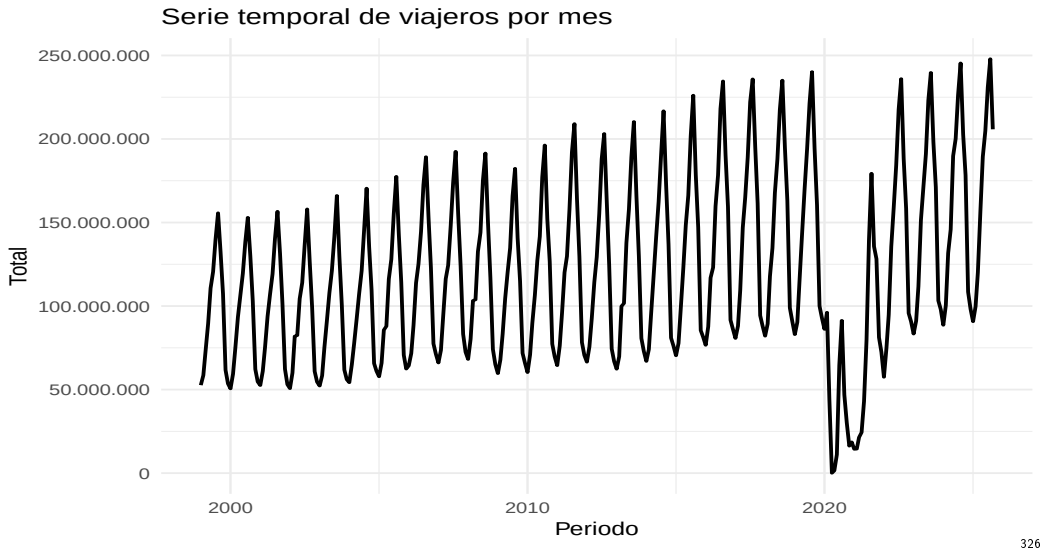
una distribución relativamente equilibrada aunque con cierta asimetría hacia los valores altos.

La desviación típica, de unos 52 millones, junto con una varianza elevada, refleja una alta variabilidad en el volumen de turismo entre periodos.

El valor mínimo (0) corresponde al impacto de la pandemia de la COVID-19, mientras que el máximo (247,9 millones) evidencia los picos alcanzados en los años de mayor actividad turística.

Para ilustrar la evolución del turismo en el conjunto nacional, se elaboró una serie temporal que permite observar las principales tendencias y patrones estacionales.

```
ggplot(nacional_mensual, aes(Periodo, Total)) +
  geom_line(linewidth = 0.8) +
  scale_y_continuous(labels = label_comma(big.mark = ".", decimal.mark = ","))+
  labs(title = "Serie temporal de viajeros por mes",
       x = "Periodo", y = "Total") +
  theme_minimal()
```

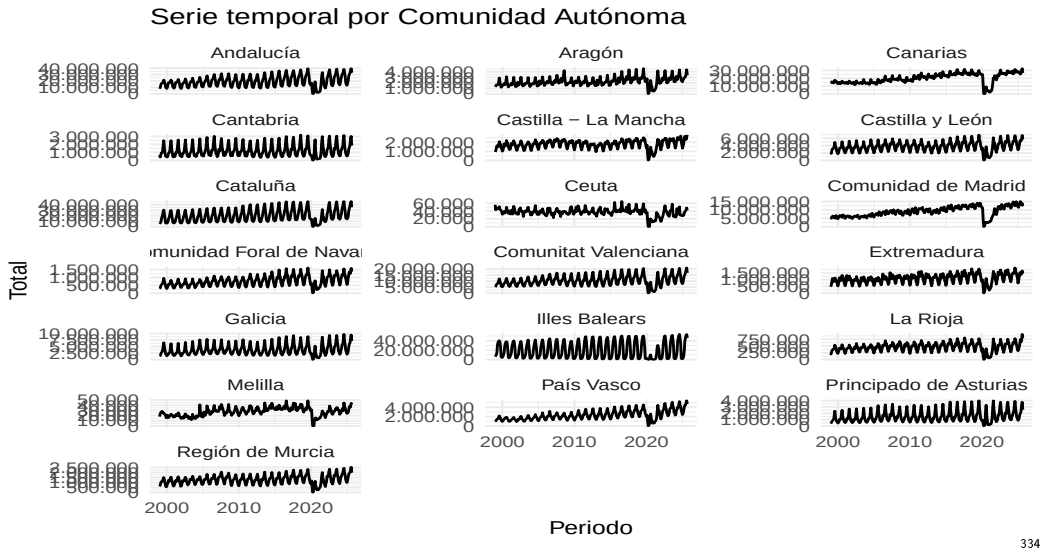


La gráfica muestra una tendencia general al alza con una clara estacionalidad anual, marcada por picos recurrentes en los meses de mayor actividad turística. También se identifican valores atípicos durante el periodo de la pandemia de la COVID-19, que reflejan la brusca caída del sector en esos años.

A continuación, se representa la serie temporal de cada comunidad y ciudad autónoma para analizar la evolución del turismo y comparar sus comportamientos a lo largo del tiempo.

```
ca_mensual <- df_Turismo %>%
  group_by(CCAA, Periodo) %>%
  summarise(Total = sum(Total, na.rm = TRUE), .groups = "drop")

ggplot(ca_mensual, aes(Periodo, Total)) +
  geom_line(linewidth = 0.6) +
  facet_wrap(~ CCAA, scales = "free_y", ncol = 3) +
  scale_y_continuous(labels = label_comma(big.mark = ".", decimal.mark = ","))+
  labs(title = "Serie temporal por Comunidad Autónoma",
       x = "Periodo", y = "Total") +
  theme_minimal()
```



En general, todas las comunidades muestran una marcada estacionalidad, con picos de actividad durante los meses de verano. A largo plazo, Madrid, el País Vasco y Canarias presentan una tendencia claramente creciente, mientras que otras, como Ceuta, mantienen una evolución más estable.

Asimismo, se observa heterocedasticidad en regiones como Illes Balears, Asturias, Andalucía o la Comunitat Valenciana, donde la variabilidad aumenta con el número de turistas, a diferencia de Madrid o Canarias, donde la dispersión se mantiene más constante.

Por último, todas las comunidades registran una caída abrupta en 2020, asociada al impacto de la pandemia de la COVID-19. En el siguiente apartado se aplicará el método 3sigma estacional (por mes y comunidad autónoma) para identificar de forma formal los valores atípicos sin confundirlos con la estacionalidad natural de la serie.

##TODO METER LOS MAPAS Y EXPLICARLO Mas bien, meter el del año 2024 y explicarlo. También estaría bien meter algo más de bivariante

3.2. OCUPACIÓN

El segundo conjunto de datos corresponde a la ocupación por sectores económicos en España, obtenido del INE. Contiene el número de personas ocupadas (en miles) por comunidad o ciudad autónoma, sector y periodo trimestral.

El dataset está compuesto por 21.300 observaciones (filas) y 5 variables (columnas), cuyos nombres son claros y representativos, lo que facilita su comprensión y manipulación posterior.

Las variables incluidas son:

- **Sexo**. Variable categórica que presenta el valor “mujer”, “hombre” y “ambos sexos”.
- **Comunidades.y.Ciudades.Autónomas**. Variable categórica que incluye las comunidades Españolas y el total nacional.
- **Sector.económico**. Variable categórica que clasifica la actividad económica en agricultura, industria, construcción o servicios.
- **Periodo**. Variable tipo fecha que muestra el año y trimestre.
- **Total**. Variable numérica que muestra las personas ocupadas (en miles).

Como primer paso, se utilizó la función `glimpse()` para revisar la estructura del dataset. Se comprobó que todas las columnas se importaron como tipo carácter, por lo que será necesario ajustar los tipos de datos en los siguientes pasos.

3.2.1. Limpieza

Para limpiar y preparar la base de datos de ocupación, ha sido necesario realizar los siguientes pasos:

1. Eliminar la columna "Sexo".
2. Eliminar números y carácteres especiales de los nombres de las comunidades. Además de filtrar y eliminar el valor "Total nacional" de la variable CCAA, por ser redundante respecto al resto de comunidades y ciudades autónomas. Posteriormente, esta variable se convierte a tipo factor.
3. Filtrar los registros correspondientes al sector "Servicios" dentro de la columna Sector.económico. Dado que la tabla se centrará exclusivamente en este sector, la variable se elimina por resultar redundante.
4. Convertir la variable "Periodo" del formato YYYYTQ a un formato de fecha trimestral mediante la función `as.yearqtr()`.
5. Normalizar la variable "Total", eliminando la palabra "miles" y los puntos de millares, sustituyendo la coma decimal por punto, convirtiendo los valores a numéricos y multiplicándolos por 1.000 para expresarlos en personas. De este modo, se mantiene la coherencia de escala con la base de datos de turismo.

```
df_Ocupacion_clean <-  df_Ocupacion %>%
  #1
  filter(Sexo == "Ambos sexos") %>%
  select(-Sexo) %>%
  #2
  rename(CCAA=Comunidades.y.Ciudades.Autónomas) %>%
  mutate(CCAA = str_replace(CCAA, "^[0-9]+\\s+", ""),
  CCAA = str_replace(CCAA,
                     "^\\s*([^,]+?)\\s*,\\s*([^,]+?)\\s*$", "\\2 \\1")) %>%
  filter(CCAA != "Total Nacional") %>%
  mutate(CCAA=as.factor(CCAA)) %>%
  #3
  filter(Sector.económico=='Servicios') %>%
  select(-Sector.económico) %>%
  #4
  mutate(Periodo = str_replace(Periodo, "T", "-"),
    Periodo = as.yearqtr(Periodo)) %>%
  #5
  mutate(
    Total=str_replace_all(Total, '\\.',""),
    Total = str_replace(Total, ",", "."),
    Total = as.numeric(Total)*1000
  ) %>%
  arrange(CCAA, Periodo)
```

3.2.2. Análisis

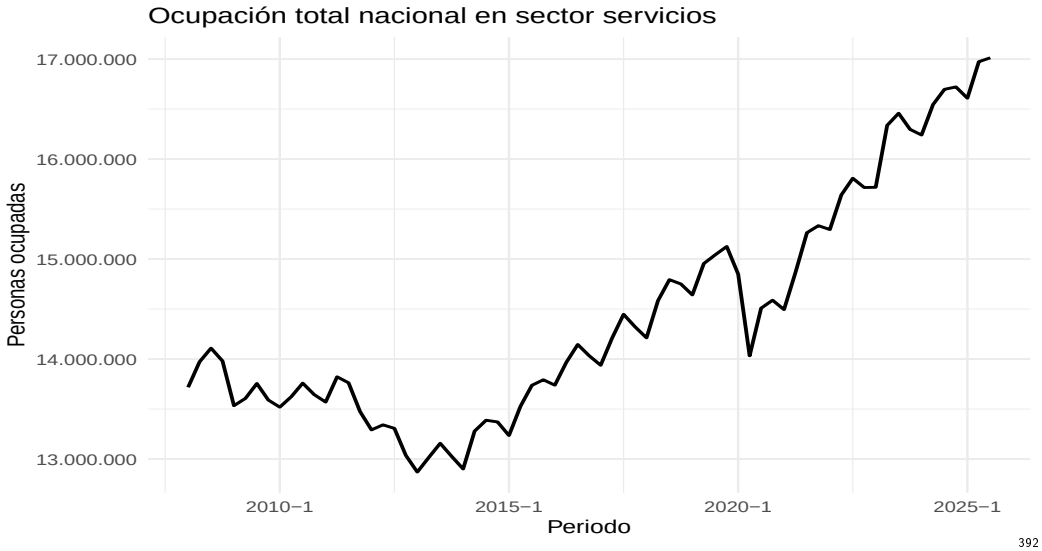
A continuación, se representa la **evolución trimestral del empleo nacional en el sector servicios**. Para ello, se agrupan los datos por periodo y se calcula el total de personas ocupadas, generando una serie temporal mediante un gráfico de líneas.

```
# Total nacional por trimestre
ocup_nacional <- df_Ocupacion_clean %>%
  group_by(Periodo) %>%
```

```
  summarise(Total = sum(Total, na.rm = TRUE), .groups = "drop")

ggplot(ocup_nacional, aes(Periodo, Total)) +
  geom_line(linewidth = 0.8) +
  scale_y_continuous(labels = label_comma(big.mark = ".", decimal.mark = ","))+
  labs(title =
         "Ocupación total nacional en sector servicios",
       x = "Periodo", y = "Personas ocupadas") +
  theme_minimal()
```

```
## Warning: The ‘trans‘ argument of ‘continuous_scale()‘ is deprecated as of ggplo
## i Please use the ‘transform‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Se observa una caída entre 2010 y 2014, posiblemente asociada a los efectos de la crisis de 2008, seguida de una recuperación sostenida hasta 2019.

En 2020 se produce un descenso notable debido al impacto de la pandemia, tras lo cual la ocupación retoma una tendencia creciente, alcanzando niveles superiores a los anteriores.

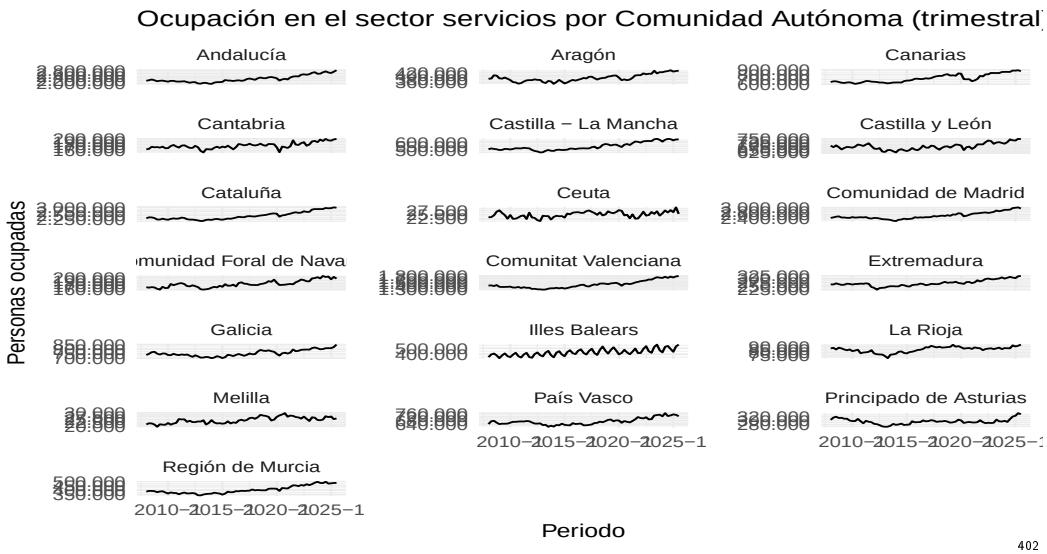
Además, se aprecia una estacionalidad clara, con aumentos recurrentes en determinados trimestres, probablemente vinculados a la temporada turística.

Además, también se analiza la **evolución trimestral del empleo en el sector servicios por Comunidad Autónoma**, agrupando los datos por región y periodo. El gráfico permite comparar la dinámica laboral entre territorios a lo largo del tiempo.

```
ocup_ccaa <- df_Ocupacion_clean %>%
  group_by(CCAA, Periodo) %>%
  summarise(Total = sum(Total, na.rm = TRUE), .groups = "drop")

ggplot(ocup_ccaa, aes(Periodo, Total)) +
  geom_line(linewidth = 0.5) +
  facet_wrap(~ CCAA, scales = "free_y", ncol = 3) +
  scale_y_continuous(labels = label_comma(big.mark = ".", decimal.mark = ","))+
  labs(title =
         "Ocupación en el sector servicios por Comunidad Autónoma (trimestral)",
       x = "Periodo", y = "Personas ocupadas") +
  theme_minimal() +
```

```
  theme(panel.spacing = unit(1, "lines"),
        strip.text = element_text(size = 9))
```



Se observa una tendencia al alza más pronunciada en comunidades como Andalucía, Canarias, Madrid y Cataluña, principales motores del sector servicios. Destaca una fuerte estacionalidad en Illes Balears, asociada a la actividad turística, mientras que en la mayoría de regiones la evolución es más estable y con un crecimiento moderado tras la pandemia.

3.2.3. Detección de outliers

Aplicando los tres métodos de detección de valores atípicos, al igual que en la base de datos del turismo, se obtiene el siguiente número de outliers identificados por cada método:

```
## # A tibble: 19 x 4
##    CCAA                       n_hampel n_boxplot n_sigma3
##    <fct>                         <int>     <int>    <int>
##  1 Extremadura                       6         8        0
##  2 Principado de Asturias            5         7        0
##  3 Castilla - La Mancha              3         0        0
##  4 La Rioja                          3         1        0
##  5 Comunidad de Madrid               2         1        0
##  6 Andalucía                         1         3        0
##  7 Cataluña                          1         0        0
##  8 Galicia                           1         0        0
##  9 País Vasco                        1         3        0
## 10 Aragón                            0         0        0
## 11 Canarias                          0         0        0
## 12 Cantabria                         0         3        0
## 13 Castilla y León                   0         0        0
## 14 Ceuta                             0         0        0
## 15 Comunidad Foral de Navarra        0         0        0
## 16 Comunitat Valenciana              0         5        0
## 17 Illes Balears                     0         0        0
## 18 Melilla                           0         0        0
## 19 Región de Murcia                  0         0        0
```

El número de valores atípicos detectados en la serie de ocupación es reducido y no presenta un patrón sistemático entre comunidades autónomas. En la mayoría de los

casos, estas observaciones reflejan variaciones reales del empleo, asociadas al crecimiento estructural del mercado laboral o a episodios económicos concretos, como la crisis de 2008.

A diferencia de lo ocurrido en la base de datos turística —donde los outliers correspondientes al periodo de la pandemia de la COVID-19 se eliminaron por representar una disrupción excepcional—, en este caso se considera que las fluctuaciones detectadas forman parte del comportamiento normal de la serie.

Por tanto, no se eliminarán las observaciones atípicas y se mantendrán en el análisis, tomando como referencia el método de detección 3-sigma, que ofrece una estimación más estable y coherente con la naturaleza de estos datos.

4. ANÁLISIS BIVARIANTE

##TODO

Base de datos actualizada con los valores imputados

```
AnalisisBox <- readRDS("../AnalisisBox.rds")

df <- data.frame(matrix(nrow = 0, ncol = 12))
ccaa <- levels(df_Turismo$CCAA)
j <- 1

for (x in AnalisisBox){
  if (inherits(x, "forecast")){
    df[j, ] <- x$mean
    rownames(df)[j] <- ccaa[[j]]
    j <- j + 1
  }
}
```

```
# Copia del data frame original
df_Turismo_corregido <- df_Turismo

# Ordenar por CCAA y Periodo para tener los meses en orden dentro de cada comunida
df_Turismo_corregido <- df_Turismo_corregido %>%
  arrange(CCAA, Periodo)

# Definir rango COVID (12 meses: 2020-03 a 2021-02)
inicio_covid <- as.Date("2020-03-01")
fin_covid    <- as.Date("2021-02-01")

# Aseguramos que df es data.frame y tiene dimensiones


# Número de meses que tenemos predichos (columnas de df)
n_meses <- ncol(df)

# Comunidades en el mismo orden para df
ccaa_niveles <- levels(df_Turismo_corregido$CCAA)

# Opcional pero útil: asignar nombres de fila a df según las CCAA
if (nrow(df) == length(ccaa_niveles)) {
  rownames(df) <- ccaa_niveles
}

# Bucle: para cada CCAA, sustituir los 12 meses de COVID por los del df
```

```r
for (ccaa_actual in ccaa_niveles) {

  # índices de las filas de esa CCAA en el periodo COVID
  idx <- which(
    df_Turismo_corregido$CCAA   == ccaa_actual &
    df_Turismo_corregido$Periodo >= inicio_covid &
    df_Turismo_corregido$Periodo <= fin_covid
  )

  # si hay tantas filas como meses predichos, sustituimos
  if (length(idx) == n_meses && ccaa_actual %in% rownames(df)) {
    df_Turismo_corregido$Total[idx] <- as.numeric(df[ccaa_actual, 1:n_meses])
  }
}
```

En esta sección se abordará el análisis bivariante, con el objetivo de relacionar las bases de datos de turismo y ocupación para estudiar posibles vínculos entre ambas variables.

Antes de combinar los conjuntos de datos, será necesario estandarizar ciertas variables con el fin de garantizar una fusión correcta (merge) y obtener un marco de datos coherente para el análisis conjunto.

Así pues, se convierte la variable Periodo al formato año-trimestre, luego se agrupan los datos por comunidad y periodo, y se calcula el total de turistas por trimestre y CCAA. Finalmente, se ordenan los registros por comunidad y periodo para facilitar el merge posterior.

```r
turismo_to_merge <- df_Turismo_corregido %>%
  mutate(Periodo = as.yearqtr(Periodo)) %>%
  group_by(CCAA, Periodo) %>%
  summarise(Total_turismo = sum(Total, na.rm = TRUE), .groups = "drop") %>%
  arrange(CCAA, Periodo)
```

Mediante la función `merge()`, se unen ambos conjuntos de datos utilizando como claves comunes las variables CCAA y Periodo.

Antes de la unión, en la base de ocupación se renombra la variable Total como Numero_ocupados para distinguirla del total de turistas y evitar ambigüedades en el nuevo dataset combinado (df_Turismo_Ocupacion).

```r
df_Turismo_Ocupacion <- merge(
  x = turismo_to_merge,
  y = (df_Ocupacion_clean %>%
    rename(Numero_ocupados = Total)),
  by.x = c("CCAA", "Periodo"),
  by.y = c("CCAA", "Periodo"))
```

Finalmente, una vez preparadas y combinadas las bases de datos, se guardan los objetos resultantes en formato `.RData`.

Esto permite conservar los datos limpios y listos para su análisis, facilitando su carga posterior sin necesidad de repetir el proceso de limpieza y fusión.

Correlaciones de Pearson

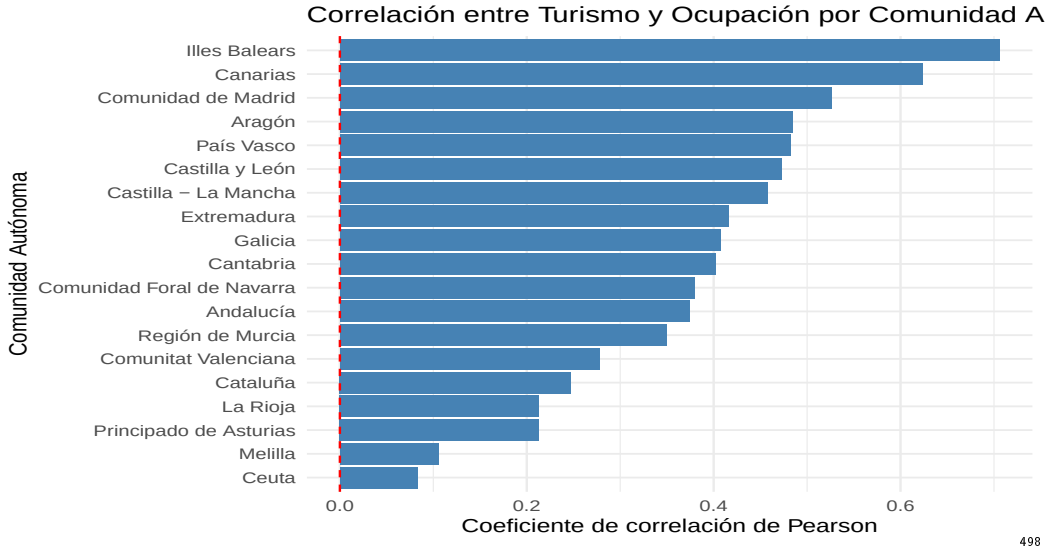
Como primer paso, se calcula la correlación de Pearson entre el número de turistas y el empleo en el sector servicios para cada Comunidad Autónoma.

El objetivo es medir la intensidad y dirección de la relación lineal entre ambas variables, obteniendo tanto el coeficiente de correlación como su p-valor para evaluar la significatividad estadística.

Además, se incorpora un indicador de nivel de significación ($p < 0.05$), que permite identificar en qué comunidades la relación es estadísticamente relevante.

Los resultados se ordenan de mayor a menor correlación, destacando aquellas regiones donde el vínculo entre turismo y empleo es más fuerte y significativo.

```
## # A tibble: 19 x 4
##    CCAA                          cor  p_value sign
##    <fct>                       <dbl>    <dbl> <chr>
##  1 Illes Balears               0.706  6.24e-12 Sí
##  2 Canarias                    0.624  6.30e- 9 Sí
##  3 Comunidad de Madrid         0.526  2.45e- 6 Sí
##  4 Aragón                      0.484  1.86e- 5 Sí
##  5 País Vasco                  0.482  2.06e- 5 Sí
##  6 Castilla y León             0.473  3.17e- 5 Sí
##  7 Castilla - La Mancha        0.458  5.99e- 5 Sí
##  8 Extremadura                 0.416  3.05e- 4 Sí
##  9 Galicia                     0.407  4.29e- 4 Sí
## 10 Cantabria                   0.402  5.07e- 4 Sí
## 11 Comunidad Foral de Navarra  0.380  1.09e- 3 Sí
## 12 Andalucía                   0.374  1.30e- 3 Sí
## 13 Región de Murcia            0.350  2.80e- 3 Sí
## 14 Comunitat Valenciana        0.278  1.88e- 2 Sí
## 15 Cataluña                    0.247  3.76e- 2 Sí
## 16 La Rioja                    0.213  7.44e- 2 No
## 17 Principado de Asturias      0.213  7.45e- 2 No
## 18 Melilla                     0.105  3.82e- 1 No
## 19 Ceuta                       0.0834 4.89e- 1 No
```



Se observa que Illes Balears presenta la mayor correlación positiva ($r = 0.68$) entre turismo y empleo en el sector servicios, reflejando su fuerte dependencia del turismo estival. En el extremo opuesto, Melilla muestra una correlación negativa ($r = -0.24$), probablemente por una menor relación directa entre la llegada de turistas y el empleo local.

En general, las correlaciones son positivas pero de magnitud moderada, indicando que ambas variables tienden a evolucionar conjuntamente, aunque con distinta intensidad según la región.

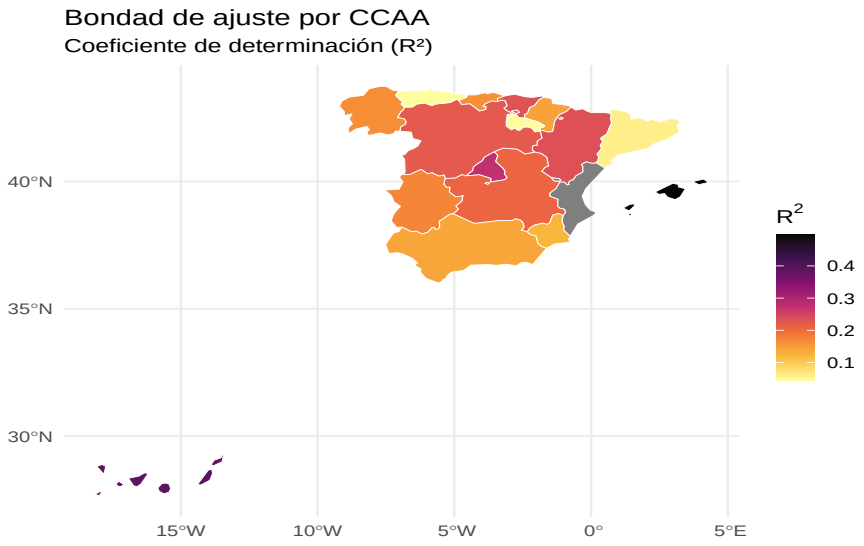
En la mayoría de comunidades, la relación es estadísticamente significativa ($p < 0.05$), mientras que en Ceuta, La Rioja, Cataluña y Asturias no se observa significatividad, posiblemente por el menor peso del turismo en su economía o por una mayor volatilidad en los datos.

Por último, cabe señalar que la correlación no implica causalidad: los resultados reflejan asociación, pero no necesariamente una relación directa de causa y efecto.

Análisis de regresión lineal simple

Además, modelamos para cada comunidad una regresión simple con el objetivo de conocer la **bondad de ajuste** (`R^2`), lo cual permitea valorar qué proporción de la variabilidad en el empleo es explicada por el turismo.

```
## # A tibble: 19 x 2
## # Rowwise:
##    CCAA                          R2
##    <fct>                       <dbl>
##  1 Andalucía                 0.140
##  2 Aragón                    0.235
##  3 Canarias                  0.389
##  4 Cantabria                 0.162
##  5 Castilla - La Mancha      0.209
##  6 Castilla y León           0.223
##  7 Cataluña                  0.0612
##  8 Ceuta                     0.00695
##  9 Comunidad de Madrid       0.277
## 10 Comunidad Foral de Navarra 0.144
## 11 Comunitat Valenciana      0.0774
## 12 Extremadura               0.173
## 13 Galicia                   0.166
## 14 Illes Balears             0.498
## 15 La Rioja                  0.0454
## 16 Melilla                   0.0111
## 17 País Vasco                0.233
## 18 Principado de Asturias    0.0454
## 19 Región de Murcia          0.122
```



En general, se observa que los valores son moderados o bajos, lo que indica que, aunque existe cierta relación entre ambas variables, el turismo no explica completamente la evolución del empleo, y probablemente influyen otros factores económicos o estructurales.

Las comunidades con mayor ajuste del modelo son Aragón ($R^2 = 0.25$), Castilla y León ($R^2 = 0.24$), Cantabria ($R^2 = 0.22$) y Canarias ($R^2 = 0.22$), donde alrededor de una cuarta parte de la variabilidad en el empleo se asocia con los cambios en el turismo.

Por el contrario, en Ceuta ($R^2 \approx 0$) y Cataluña ($Rapprox0.05$) el modelo lineal apenas explica la variación en el empleo, lo que sugiere una relación débil o inexistente entre ambas variables en estas regiones.

A partir del modelo de regresión lineal calculado anteriormente, se estima la **elasticidad** del empleo respecto al turismo, utilizando la pendiente de la recta de regresión como medida del impacto del turismo sobre el empleo en cada comunidad autónoma.

```
## # A tibble: 19 x 3
##    CCAA                        elasticidad int
##    <fct>                             <dbl> <chr>
##  1 Canarias                          0.396  Inelástico
##  2 Castilla - La Mancha              0.211  Inelástico
##  3 Comunidad de Madrid               0.207  Inelástico
##  4 Extremadura                       0.156  Inelástico
##  5 Región de Murcia                  0.130  Inelástico
##  6 Illes Balears                     0.117  Inelástico
##  7 Aragón                            0.109  Inelástico
##  8 Andalucía                         0.0949 Inelástico
##  9 Comunitat Valenciana              0.0825 Inelástico
## 10 País Vasco                        0.0816 Inelástico
## 11 Comunidad Foral de Navarra        0.0777 Inelástico
## 12 Castilla y León                   0.0752 Inelástico
## 13 Melilla                           0.0563 Inelástico
## 14 Cataluña                          0.0528 Inelástico
## 15 La Rioja                          0.0452 Inelástico
## 16 Cantabria                         0.0418 Inelástico
## 17 Galicia                           0.0407 Inelástico
## 18 Ceuta                             0.0295 Inelástico
## 19 Principado de Asturias            0.0189 Inelástico
```

Los valores de elasticidad, comprendidos entre -0.09 y 0.19, muestran que en todas las comunidades autónomas la relación entre turismo y empleo es inelástica (ya que <1), es decir, el empleo en el sector servicios apenas varía ante los cambios en la actividad turística. Esto sugiere que un aumento del turismo no se traduce proporcionalmente en un mayor nivel de empleo.

Esta baja respuesta puede deberse a que parte del empleo es estructural o fijo, a la existencia de desfases temporales entre el incremento del turismo y la contratación de personal, o a que las variables se miden en frecuencias distintas (mensual para el turismo y trimestral para el empleo), lo que suaviza la relación observada.

En regiones como Canarias (≈ 0.19), donde la elasticidad es algo mayor, el empleo muestra una mayor sensibilidad a las variaciones del turismo, coherente con su elevada dependencia económica de esta actividad.

Aunque las correlaciones y la bondad de ajuste indicaban una relación fuerte y significativa entre turismo y empleo, las elasticidades calculadas son bajas. Esto no representa una contradicción, sino una diferencia en la interpretación: mientras la correlación mide la fuerza de la relación, la elasticidad cuantifica su intensidad relativa.

En otras palabras, aunque el turismo y el empleo evolucionan de forma coordinada, el impacto porcentual del turismo sobre el empleo es reducido, lo que sugiere que otros factores —como la estacionalidad o la estabilidad estructural del mercado laboral— también influyen en la evolución del empleo.

5. CONCLUSIONES