

Trabajo Opcional Edig

Mihai Cristian Mihalache Farcas y Rubén Tormo Piles

2025-11-10

Contents

1. Introducción	3
2. Datos, de donde los hemos sacado y que representan.	3
2.1. Pernoctaciones turísticas:	3
2.2. Plazas de alojamiento	3
2.3. Índice de Producción Industrial (IPI)	4
2.4. Índice de Precios de Consumo (IPC)	4
2.5. Datos de precios: ADR	5
2.6. Construcción del panel y variable temporal	5
3. Análisis descriptivo	6
3.1. Estadísticos descriptivos	6
3.2. Gráficos descriptivos	8
3.2.1. Evolución agregada de las pernoctaciones	8
3.2.2. Evolución de las pernoctaciones por CCAA	9
3.2.3. Distribución de pernoctaciones por CCAA	10
3.3. Matriz de correlaciones	12
4. Modelo económico de datos de panel	13
4.1. Formulación de la pregunta de interés	13
4.2. Modelo teórico	14
4.3. Especificación econométrica	14
4.3.1. Variable dependiente y explicativas	14
4.3.2. Transformaciones (logs, ADR real, etc.)	14

5. Estimación de modelos de panel	14
5.1. Modelo pooled OLS	14
5.2. Modelo de efectos fijos	16
5.3. Modelo de efectos aleatorios	17
5.4. Comparación de modelos	18
5.4.1. Test F (EF vs pooled)	18
5.4.2. Test de Hausman (EF vs EA)	19
5.5. Diagnóstico del modelo de efectos fijos	19
5.5.1. Errores robustos a heteroscedasticidad y autocorrelación	19
5.5.2. Corrección con errores robustos	20
6. Estimador de diferencias en diferencias	20
7. Interpretación de resultados de panel	22
7.1. Efectos de plazas, ADR real e IPI	22
7.2. Discusión económica de los coeficientes	22
8. Análisis de series temporales	22
8.1. Selección de la serie (por ejemplo, pernoctaciones totales)	23
8.2. Exploración y descomposición	24
8.2.1. Tendencia, ciclo y estacionalidad	25
8.3. Transformaciones y estacionariedad	26
8.4. Modelización con <code>forecast</code>	26
8.4.1. Modelos ARIMA/ETS	27
8.4.2. Selección de modelo	29
9. Evaluación de pronósticos	30
9.1. Diseño de la ventana deslizante	30
9.2. Métricas de error (MAE, RMSE, etc.)	30
9.3. Comparación de modelos de pronóstico	30
10. Pronóstico final e interpretación	31
10.1. Gráficos de pronóstico con bandas de confianza	31
10.2. Lectura económica de los resultados	32
11. Conclusiones	32
11.1. Resumen de hallazgos del panel	32
11.2. Resumen de resultados de series temporales	33
11.3. Limitaciones y posibles extensiones	33

1. Introducción

El objetivo de este trabajo es **analizar empíricamente el comportamiento del turismo en las comunidades autónomas españolas** utilizando técnicas de **datos de panel y series temporales**. Se emplean datos mensuales de pernoctaciones, plazas de alojamiento, precio medio por noche (ADR) e índice de producción industrial (IPI), construyendo un panel regional con frecuencia inferior al año.

El análisis se estructura en dos grandes bloques. En primer lugar, se desarrolla un **modelo de datos de panel** que relaciona las pernoctaciones con sus principales determinantes (capacidad, precios y ciclo económico), siguiendo los pasos clásicos del análisis económico empírico: formulación de la pregunta de interés, construcción del modelo teórico, estimación de diferentes especificaciones (pooled, efectos fijos y efectos aleatorios) y elección del modelo más adecuado mediante contrastes estadísticos.

En segundo lugar, se selecciona una de las variables del panel para realizar un **ejercicio de pronóstico de series temporales**. Para ello se trabaja con el paquete **forecast**, se aplican transformaciones y técnicas de descomposición, se estiman modelos ARIMA/ETS y se evalúa la capacidad predictiva mediante una **ventana deslizante**, comparando varios modelos en términos de error de predicción.

A lo largo del informe se combinan resultados numéricos y gráficos con una interpretación económica cuidadosa, de modo que las conclusiones estén respaldadas tanto por la evidencia estadística como por el contexto del sector turístico español.

2. Datos, de donde los hemos sacado y que representan.

2.1. Pernoctaciones turísticas:

ccaa	anio	mes	pernoctaciones
01 Andalucía	2025	10	5250858
01 Andalucía	2025	9	6176273
01 Andalucía	2025	8	7586225
01 Andalucía	2025	7	6857986
01 Andalucía	2025	6	6009338
01 Andalucía	2025	5	5582983

La variable principal de estudio son las **pernoctaciones mensuales en establecimientos turísticos por comunidad autónoma**, obtenidas a partir de las estadísticas oficiales de ocupación publicadas por el Instituto Nacional de Estadística (INE). Esta información recoge el número total de noches que los viajeros pasan alojados en hoteles y otros tipos de alojamiento reglado, lo que constituye un indicador directo del nivel de actividad turística en cada región y mes.

En la base de datos, estas observaciones se almacenan en la columna **pernoctaciones**, junto con los identificadores de comunidad autónoma (**ccaa**), año (**anio**) y mes (**mes**), permitiendo construir un panel mensual para el periodo analizado.

2.2. Plazas de alojamiento

ccaa	anio	mes	plazas
01 Andalucía	2025	10	282892
01 Andalucía	2025	9	321850
01 Andalucía	2025	8	324083

ccaa	anio	mes	plazas
01 Andalucía	2025	7	323748
01 Andalucía	2025	6	316048
01 Andalucía	2025	5	304525

Como medida de la **capacidad de oferta turística**, se utilizan las **plazas disponibles en establecimientos turísticos** por comunidad autónoma y mes, también procedentes de las encuestas de ocupación del INE. Esta variable refleja el número máximo de pernoctaciones que pueden atenderse en cada destino y periodo, y permite capturar diferencias estructurales en el tamaño y grado de desarrollo del sector turístico regional.

Esta información se recoge en la columna **plazas**, asociada igualmente a los identificadores temporales y regionales, lo que facilita su uso como variable explicativa en el modelo de datos de panel.

2.3. Índice de Producción Industrial (IPI)

ccaa	anio	mes	ipi
01 Andalucía	2025	10	120.549
01 Andalucía	2025	9	113.225
01 Andalucía	2025	8	96.561
01 Andalucía	2025	7	118.808
01 Andalucía	2025	6	113.461
01 Andalucía	2025	5	108.785

Como indicador del **ciclo económico regional**, se incorpora el **Índice de Producción Industrial (IPI)** por comunidad autónoma, normalmente publicado con base 100 por el INE. El IPI resume la evolución de la actividad productiva, y se utiliza aquí como variable de control para capturar el efecto del entorno macroeconómico sobre la demanda turística.

En la base, el IPI se almacena en la columna **ipi**. Al estar medido en índice, valores superiores a 100 indican un nivel de producción por encima del año base, mientras que valores inferiores reflejan fases de desaceleración.

2.4. Índice de Precios de Consumo (IPC)

ccaa	2025M1	2025M0	2024M12							
01 Andalucía	120.163	119.106	119.309	119.321	119.641	118.997	118.800	118.076	117.919	117.486
02 Aragón	119.472	118.557	118.912	119.025	119.022	118.200	118.218	117.510	117.360	116.915
03 Asturias, Principado de	119.205	118.655	119.688	119.258	119.071	117.974	118.191	117.246	117.269	116.807
04 Balears, Illes	120.201	119.683	120.410	120.350	119.989	118.830	118.786	117.717	117.608	116.989
05 Canarias	118.784	117.759	118.537	118.326	118.238	117.678	117.842	116.940	116.972	116.443
06 Cantabria	119.125	118.101	118.985	118.575	118.622	117.559	117.703	116.854	116.945	116.672

El **Índice de Precios de Consumo (IPC)** se emplea exclusivamente como deflactor del ADR. El IPC se toma de las estadísticas oficiales y se mide como un índice con base 100 por comunidad y mes.

2.5. Datos de precios: ADR

ccaa	anio	mes	adr
01 Andalucía	2025	10	84.45815
01 Andalucía	2025	9	94.89086
01 Andalucía	2025	8	121.39995
01 Andalucía	2025	7	110.77865
01 Andalucía	2025	6	93.45333
01 Andalucía	2025	5	86.10067

Como medida del **precio medio por noche**, se utiliza el *Average Daily Rate* (ADR) obtenido a partir de las estadísticas de ocupación hotelera. El ADR recoge la facturación media por habitación ocupada y se interpreta como el precio nominal que pagan los turistas por pernoctar en cada comunidad autónoma y mes. En la base, este valor se almacena en la columna **adr**.

Dado que el ADR está expresado en términos nominales, se construye una versión **deflactada** utilizando el Índice de Precios de Consumo (IPC) regional:

$$ADR_{real_{it}} = \frac{ADR_{it}}{IPC_{it}/100}$$

De este modo se eliminan las variaciones debidas al nivel general de precios y se obtiene una medida del coste real del alojamiento turístico a lo largo del tiempo que permite analizar la evolución del precio del alojamiento descontando el efecto de la inflación y hacer comparaciones temporales más significativas.

2.6. Construcción del panel y variable temporal

```
knitr::kable(head(df[df$ccaa == '18 Ceuta', ]))
```

ccaa	anio	mes	pernoctaciones	plazas	adr	ipi	fecha_panel
18 Ceuta	2025	10	12397	745	58.95996	NA	2025.750
18 Ceuta	2025	9	13934	776	60.14720	NA	2025.667
18 Ceuta	2025	8	16990	788	65.63376	NA	2025.583
18 Ceuta	2025	7	14411	763	63.44245	NA	2025.500
18 Ceuta	2025	6	13385	794	60.14111	NA	2025.417
18 Ceuta	2025	5	11843	769	57.26156	NA	2025.333

```
knitr::kable(head(df[df$ccaa == '19 Melilla', ]))
```

ccaa	anio	mes	pernoctaciones	plazas	adr	ipi	fecha_panel
19 Melilla	2025	10	13285	845	62.99417	NA	2025.750
19 Melilla	2025	9	14261	845	69.18685	NA	2025.667
19 Melilla	2025	8	14936	835	66.42733	NA	2025.583
19 Melilla	2025	7	12434	845	64.81537	NA	2025.500
19 Melilla	2025	6	12671	835	69.24477	NA	2025.417
19 Melilla	2025	5	12826	835	64.78199	NA	2025.333

A partir de estas fuentes, se ha construido una base integrada `df` en formato de panel, donde cada fila corresponde a una combinación **comunidad autónoma–mes–año**. Además de las columnas de identificación (`ccaa, anio, mes`), se ha creado una variable temporal continua `fecha_panel` (año fraccionado) para facilitar la representación gráfica y el tratamiento como serie temporal agregada.

Este diseño permite combinar en un mismo conjunto de datos las dimensiones **transversal** (diferencias entre CCAA) y **temporal** (evolución mensual), cumpliendo los requisitos del trabajo para el análisis de datos de panel y series temporales.

Pero, como podemos observar, no existen datos sobre el IPI de Ceuta y Melilla en el INE. Dado que el IPI forma parte de las variables explicativas del modelo de datos de panel, mantener estas observaciones implicaría trabajar con un panel fuertemente incompleto o recurrir a imputaciones ad-hoc de difícil justificación económica. Por este motivo, se ha optado por eliminar las filas correspondientes a Ceuta y Melilla, de manera que el análisis se centra en las comunidades para las que existe información completa y comparable en todas las variables utilizadas. Esta decisión permite estimar el modelo de efectos fijos sobre un panel más coherente y evita introducir supuestos adicionales sobre la evolución no observada del IPI en estos territorios.

```
knitr::kable(head(df))
```

ccaa	anio	mes	pernoctaciones	plazas	adr	ipi	fecha_panel
01 Andalucía	2025	10	5250858	282892	84.45815	120.549	2025.750
01 Andalucía	2025	9	6176273	321850	94.89086	113.225	2025.667
01 Andalucía	2025	8	7586225	324083	121.39995	96.561	2025.583
01 Andalucía	2025	7	6857986	323748	110.77865	118.808	2025.500
01 Andalucía	2025	6	6009338	316048	93.45333	113.461	2025.417
01 Andalucía	2025	5	5582983	304525	86.10067	108.785	2025.333

Por lo tanto, `df` quedaría de esta manera.

3. Análisis descriptivo

3.1. Estadísticos descriptivos

En este apartado, presentamos los estadísticos descriptivos básicos de las principales variables del panel (pernoctaciones, plazas, adr e ipi) con el objetivo de resumir su distribución y tener una primera idea de los niveles y la dispersión del turismo y de sus posibles determinantes.

```
# Estadísticos básicos
summary(df[, c("pernoctaciones", "plazas", "adr", "ipi")])
```

```
##   pernoctaciones      plazas          adr          ipi
## Min.    :     0  Min.    :     0  Min.    :  0.00  Min.    : 55.12
## 1st Qu.: 199302  1st Qu.: 18634  1st Qu.: 50.48  1st Qu.: 93.14
## Median : 406800  Median : 34842  Median : 56.29  Median :100.83
## Mean    : 1454255  Mean    : 82848  Mean    : 62.14  Mean    :101.01
## 3rd Qu.: 1845550  3rd Qu.:115304  3rd Qu.: 71.52  3rd Qu.:108.70
## Max.    :11414856  Max.    :381103  Max.    :150.65  Max.    :163.63
##                      NA's    :34
```

En los estadísticos, podemos observar como **pernoctaciones** y **plazas** presentan medias muy superiores a sus medianas, lo que indica distribuciones claramente asimétricas hacia la derecha: algunos meses y CCAA concentran niveles de demanda y capacidad muy elevados, que empujan la media por encima del “caso típico”. Además, los valores máximos de pernoctaciones (más de 11 millones) y plazas (más de 380 mil) confirman la existencia de observaciones extremas asociadas a destinos turísticos de gran tamaño.

En contraste, **ADR** e **IPI** aparecen mucho más acotados. Sus mínimos y máximos están relativamente próximos y las medianas son muy similares a las medias, lo que sugiere una dispersión moderada y menor asimetría. Esto refleja que los precios medios por noche y la actividad industrial regional varían, pero no alcanzan los niveles de concentración observados en la demanda y la capacidad hotelera.

Por último, la existencia de algunos valores perdidos únicamente en la variable ADR indica que, aunque la información sobre pernoctaciones, plazas e IPI está completa para todos los meses y comunidades, la serie de precios no está observada en todos los casos, por lo que resulta necesario decidir explícitamente cómo tratar estas observaciones incompletas antes de proceder a la estimación econométrica.

```
df_na_adr <- df[is.na(df$adr), ]
dplyr::count(df_na_adr, anio, mes)
```

```
## # A tibble: 2 x 3
##   anio     mes     n
##   <int> <int> <int>
## 1 2020      5     17
## 2 2020      6     17
```

Los valores perdidos en ADR se concentran únicamente en dos meses muy concretos: mayo y junio de 2020, con 17 comunidades afectadas en cada uno de ellos. Dado que se trata de un episodio excepcional asociado a la pandemia y que el resto del panel está completamente observado, se ha optado por eliminar estas observaciones al construir el panel para los modelos de datos de panel. Esta decisión reduce mínimamente el tamaño muestral y evita tener que imputar precios en un periodo atípico, lo que podría distorsionar la interpretación econométrica de los resultados.

```
df <- df[!is.na(df$adr), ]
summary(df[, c("pernoctaciones", "plazas", "adr", "ipi")])
```

```
##   pernoctaciones       plazas         adr          ipi
##   Min. : 0   Min. : 0   Min. : 0.00  Min. : 55.12
##   1st Qu.: 204659  1st Qu.: 18802  1st Qu.: 50.48  1st Qu.: 93.33
##   Median : 413097  Median : 35149  Median : 56.29  Median :100.96
##   Mean   : 1469057  Mean   : 83600  Mean   : 62.14  Mean   :101.18
##   3rd Qu.: 1870252  3rd Qu.:115926  3rd Qu.: 71.52  3rd Qu.:108.78
##   Max.  :11414856  Max.  :381103  Max.  :150.65  Max.  :163.63
```

Ahora ya podemos observar como no tenemos ningún valor faltante en el panel df.

```
# Media y desviación típica por variable
desc_stats <- df |>
  summarise(
    mean_perno = mean(pernoctaciones, na.rm = TRUE),
    sd_perno   = sd(pernoctaciones, na.rm = TRUE),
    mean_plazas = mean(plazas, na.rm = TRUE),
    sd_plazas  = sd(plazas, na.rm = TRUE),
    mean_adr   = mean(adr, na.rm = TRUE),
```

```

sd_adr      = sd(adr, na.rm = TRUE),
mean_ipi   = mean(ipi, na.rm = TRUE),
sd_ipi     = sd(ipi, na.rm = TRUE)
)
desc_stats

## # A tibble: 1 x 8
##   mean_perno sd_perno mean_plazas sd_plazas mean_adr sd_adr mean_ipi sd_ipi
##       <dbl>     <dbl>      <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>
## 1    1469057.  2148967.     83600.    94234.     62.1    17.0    101.    13.2

```

Las medias y desviaciones típicas confirman la fuerte heterogeneidad entre observaciones. La **media de pernoctaciones** ronda 1,4 millones con una desviación típica superior a 2 millones, lo que implica una variabilidad muy elevada entre CCAA y meses, coherente con la existencia de destinos muy masivos frente a otros claramente secundarios.

En el caso de **plazas**, la media en torno a 83 mil y la desviación típica cercana a 94 mil apuntan a una dispersión también muy alta en la capacidad de alojamiento, lo que refuerza la idea de que la oferta está muy concentrada en pocas regiones.

Por el contrario, **ADR** e **IPI** presentan medias cercanas a 62 y 101 respectivamente, con desviaciones típicas mucho más contenidas (en torno a 16 para ADR y 13 para IPI). Esto indica que tanto los precios medios del alojamiento como el nivel de actividad industrial muestran variaciones relevantes pero relativamente moderadas en comparación con la enorme dispersión observada en la demanda y la capacidad hotelera.

3.2. Gráficos descriptivos

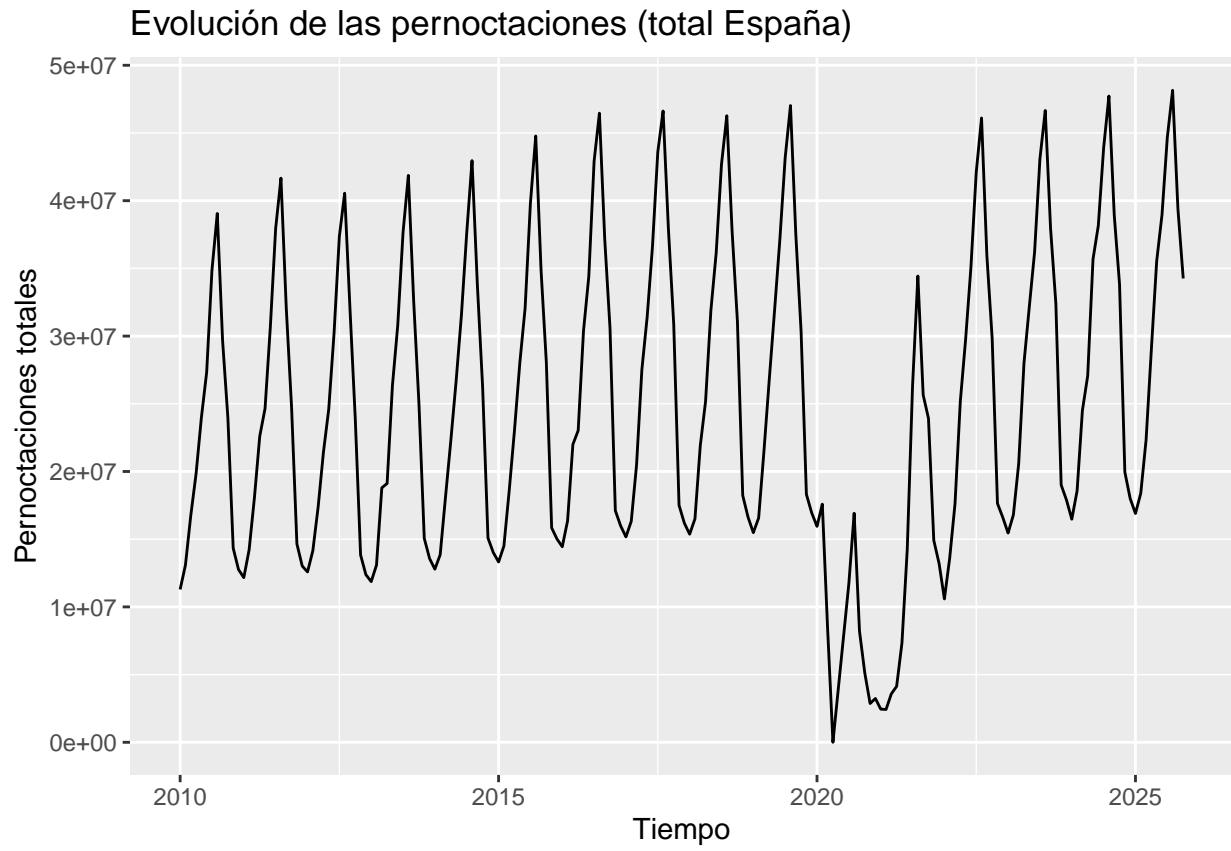
3.2.1. Evolución agregada de las pernoctaciones

Ahora, vamos a representar la suma mensual de pernoctaciones a nivel nacional con el fin de identificar tendencias de largo plazo, patrones estacionales y posibles rupturas estructurales en la serie.

```

ggplot(df, aes(x = fecha_panel, y = pernoctaciones)) +
  stat_summary(fun = sum, geom = "line") +
  labs(x = "Tiempo", y = "Pernoctaciones totales",
       title = "Evolución de las pernoctaciones (total España)")

```



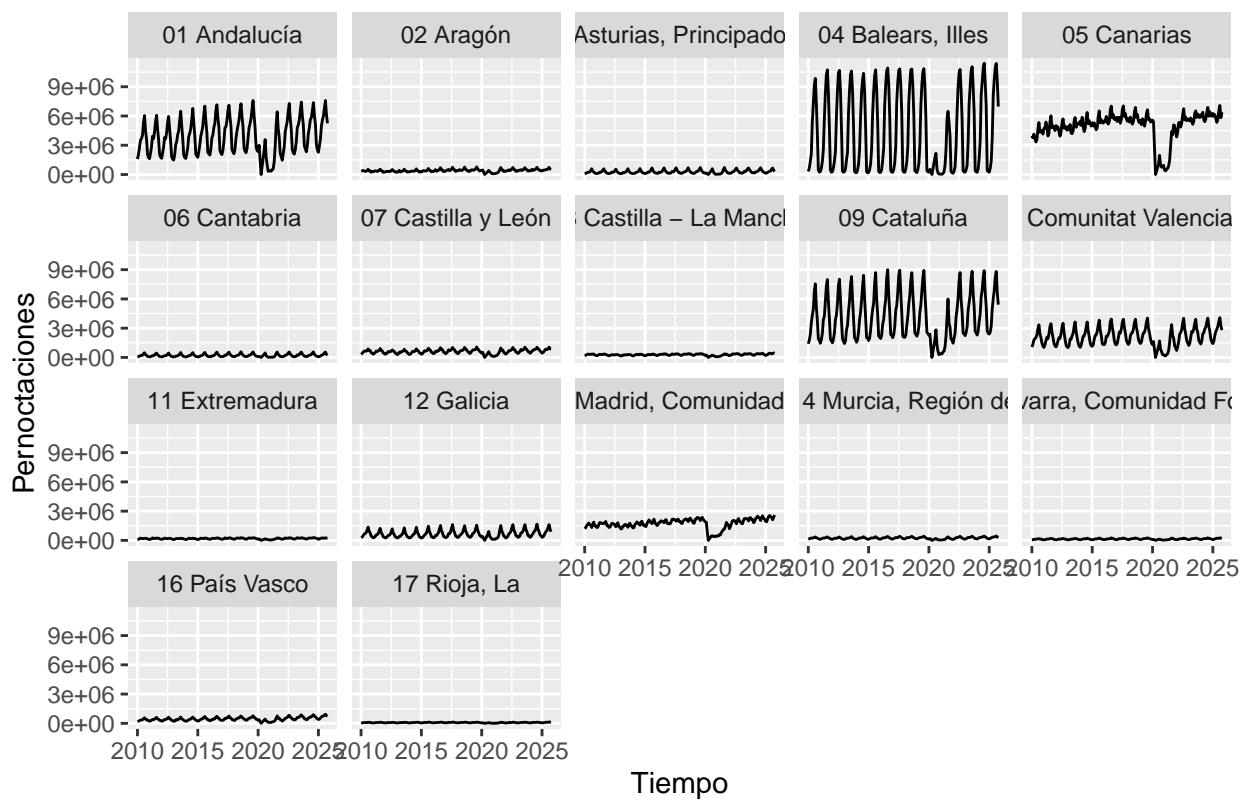
Según lo que podemos llegar a observar, la serie agregada muestra un comportamiento claramente estacional, con picos muy pronunciados en los meses de verano y mínimos en temporada baja, sobre una tendencia creciente en el período analizado. Destaca una fuerte caída en 2020, seguida de una recuperación gradual, coherente con el impacto de la pandemia sobre la actividad turística.

3.2.2. Evolución de las pernoctaciones por CCAA

A continuación se representan las pernoctaciones por comunidad autónoma mediante gráficos de líneas facetados. Esto permite comparar la dinámica temporal entre regiones y detectar diferencias estructurales en niveles y patrones de estacionalidad.

```
ggplot(df, aes(x = fecha_panel, y = pernoctaciones)) +
  geom_line() +
  facet_wrap(~ ccaa) +
  labs(x = "Tiempo", y = "Pernoctaciones",
       title = "Pernoctaciones por CCAA en el tiempo")
```

Pernoctaciones por CCAA en el tiempo



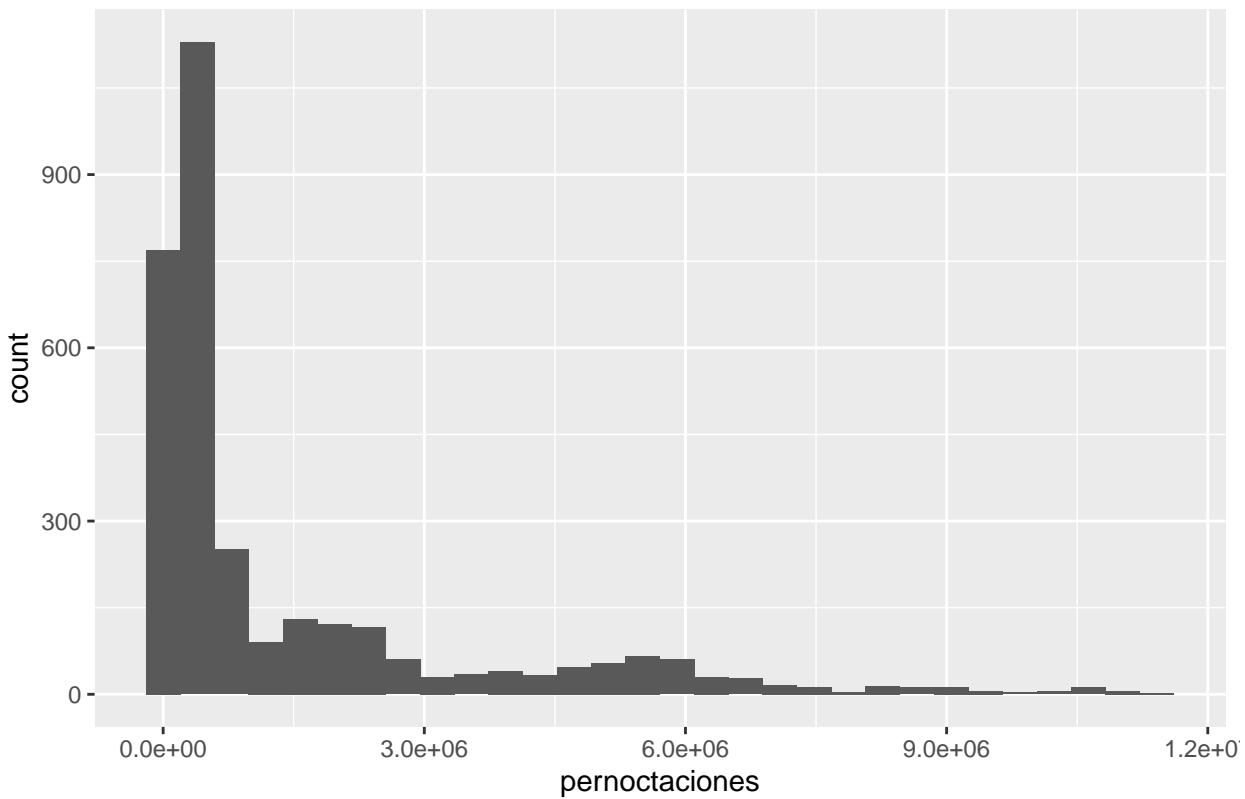
Las series regionales confirman una elevada concentración del turismo en Baleares, Canarias, Cataluña, Comunidad Valenciana y Andalucía, que presentan niveles muy superiores al resto y ciclos estacionales bien definidos. Otras regiones, como La Rioja, Navarra, Ceuta o Melilla, muestran volúmenes de pernoctaciones mucho más reducidos y, en algunos casos, mayor volatilidad relativa, lo que apunta a estructuras turísticas menos consolidadas

3.2.3. Distribución de pernoctaciones por CCAA

Utilizamos ahora diagramas de caja de pernoctaciones por comunidad autónoma para analizar la distribución de la actividad turística dentro de cada región, identificando la mediana, la dispersión y la presencia de observaciones extremas.

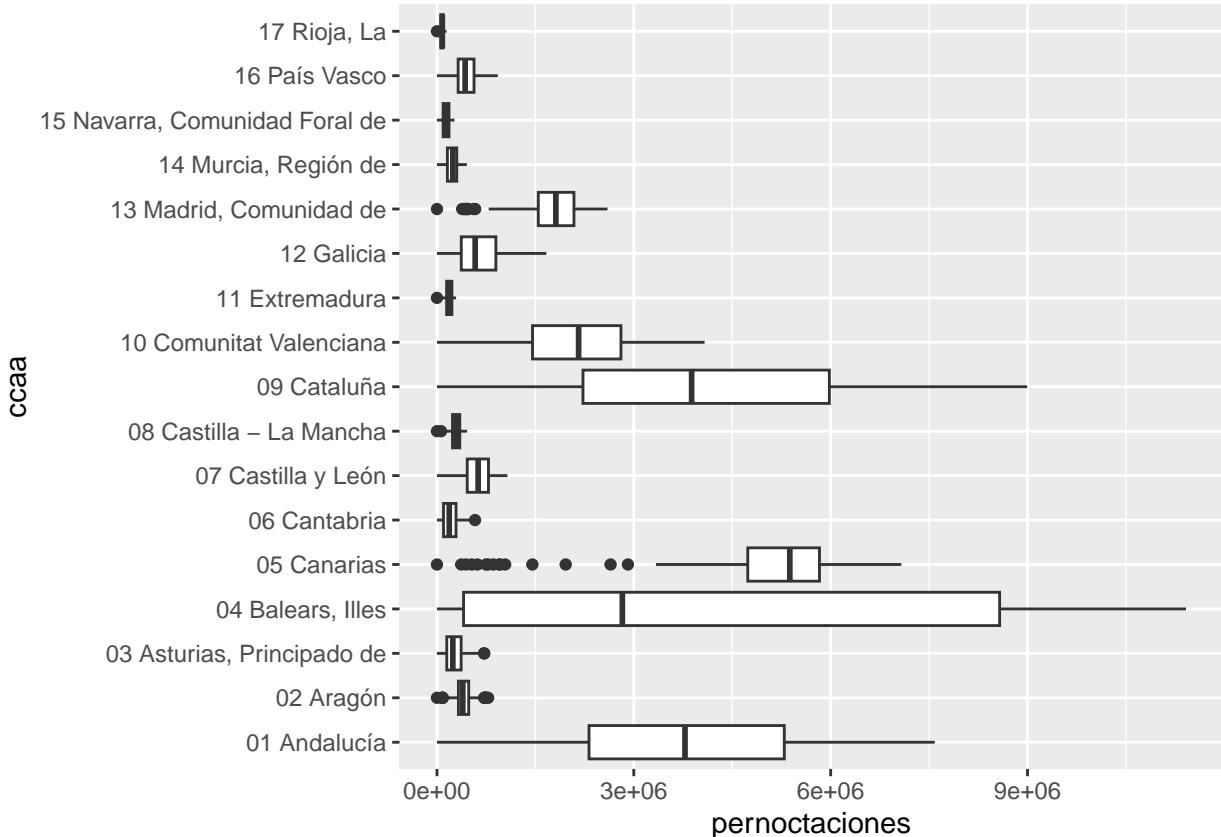
```
ggplot(df, aes(x = pernoctaciones)) +
  geom_histogram(bins = 30) +
  labs(title = "Distribución de pernoctaciones")
```

Distribución de pernoctaciones



Podemos observar una distribución de pernoctaciones muy asimétrica a la derecha ya que la mayoría de observaciones se concentran en valores bajos cercanos a 0, y a medida que aumentan las pernoctaciones la frecuencia cae rápidamente, quedando pocas observaciones con valores muy altos. Esto sugiere presencia de valores extremos y una fuerte cola derecha, por lo que transformaciones como el logaritmo podrían ser útiles para el análisis.

```
ggplot(df, aes(x = ccaa, y = pernoctaciones)) +  
  geom_boxplot() +  
  coord_flip()
```



Por otro lado, el diagrama de cajas por CCAA muestra que las pernoctaciones son muy heterogéneas entre regiones. Baleares, Cataluña y Comunitat Valenciana concentran los niveles más altos y con mayor dispersión, mientras que regiones como La Rioja, Navarra o Cantabria presentan volúmenes mucho menores y rangos intercuartílicos muy estrechos. Además, se observan valores atípicos altos en varias comunidades, lo que refuerza la idea de una distribución muy sesgada y con episodios puntuales de demanda turística extrema.

3.3. Matriz de correlaciones

Por último, se calcula la matriz de correlaciones entre pernoctaciones y las variables explicativas (plazas, adr e ipi). El objetivo es explorar las relaciones lineales bivariantes y valorar la posible existencia de multicolinealidad previa a la estimación del modelo de datos de panel.

```
vars_cor <- df |>
  select(pernoctaciones, plazas, adr, ipi)

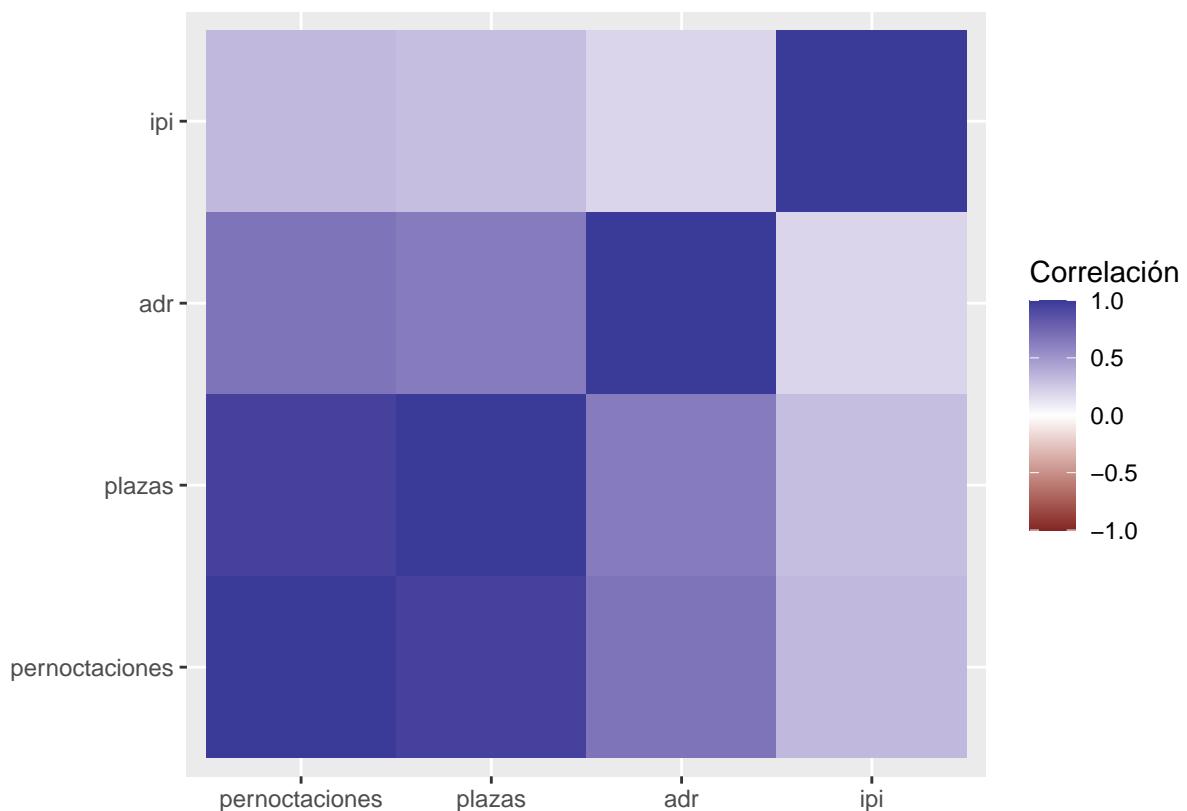
cor_matrix <- cor(vars_cor, use = "pairwise.complete.obs")
cor_matrix
```

	pernoctaciones	plazas	adr	ipi
## pernoctaciones	1.0000000	0.9572216	0.6806087	0.3328674
## plazas	0.9572216	1.0000000	0.6415886	0.3072241
## adr	0.6806087	0.6415886	1.0000000	0.1947814
## ipi	0.3328674	0.3072241	0.1947814	1.0000000

```

cor_long <- melt(cor_matrix)
ggplot(cor_long, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(limits = c(-1,1)) +
  labs(x = "", y = "", fill = "Correlación")

```



Las pernoctaciones están fuertemente correlacionadas con las plazas (correlación superior a 0,95), lo que subraya la importancia de la capacidad de alojamiento, pero también sugiere un posible problema de multicolinealidad al incluir ambas variables en la misma especificación. La correlación entre pernoctaciones y adr es positiva y de magnitud media (en torno a 0,66), indicando que los destinos más demandados tienden a asociarse con precios más elevados, posiblemente por mayor calidad o presión de demanda. La relación con el ipi es positiva pero más moderada, en torno a 0,34, lo que apunta a que el ciclo económico influye en el turismo, aunque en menor medida que la capacidad y los precios.

4. Modelo económico de datos de panel

4.1. Formulación de la pregunta de interés

La pregunta en la que nos vamos a centrar es:

¿Cómo afectan el precio real (ADR deflactado), la capacidad hotelera (plazas) y la renta disponible (IPI) a la demanda turística (pernoctaciones) en las CCAA españolas?

Se espera:

- Elasticidad-precio negativa: ($\beta_1 < 0$)
- Efecto oferta positivo: ($\beta_2 > 0$)
- Mayor demanda con mejor ciclo económico: ($\beta_3 > 0$)

4.2. Modelo teórico

$$Pernoctaciones_{it} = \beta_0 + \beta_1 \cdot ADR_{real,it} + \beta_2 \cdot Plazas_{it} + \beta_3 \cdot IPI_{it} + \sum_{m=2}^{12} \delta_m \cdot DummyMes_m + \alpha_i + \epsilon_{it}$$

Donde α_i captura heterogeneidad territorial no observada (preferencias, clima, accesibilidad).

4.3. Especificación econométrica

4.3.1. Variable dependiente y explicativas

- **Pernoctaciones totales:** miden el volumen de demanda turística agregado en cada observación, y actúan como variable dependiente del modelo.
- **ADR real:** cociente entre el ADR nominal y el índice de precios al consumo (IPC/100); recoge el precio medio por habitación ajustado por inflación, es decir, el coste “en términos reales” para el cliente.
- **Plazas estimadas:** aproximan la capacidad de oferta alojativa disponible en cada región y período, permitiendo capturar restricciones o ampliaciones de capacidad.
- **IPI general:** índice de producción industrial utilizado como proxy del ciclo económico y de la renta disponible regional, que afecta al nivel de gasto turístico.
- **Dummies mensuales:** variables ficticias que identifican cada mes del año y permiten controlar la estacionalidad fija de la demanda (picos de verano, Semana Santa, etc.).

4.3.2. Transformaciones (logs, ADR real, etc.)

- **ADR real:** el precio medio por habitación se deflacta dividiendo el ADR nominal entre el índice de precios al consumo (IPC/100), de forma que el modelo trabaja siempre con precios constantes y hace comparables períodos con distinta inflación.
- **Panel desbalanceado por CCAA:** la información se organiza como un panel regional de 17 comunidades autónomas con frecuencia mensual, pero algunas combinaciones CCAA-mes carecen de datos, por lo que el conjunto resulta ser un panel desbalanceado en lugar de completamente equilibrado.

5. Estimación de modelos de panel

La siguiente sección presenta la estimación empírica del modelo de demanda de pernoctaciones utilizando técnicas de datos de panel. Concretamente, compararemos cual es el mejor modelo que responde a nuestra pregunta de interés entre los modelos **pooled OLS**, **efectos fijos** y **efectos aleatorios**.

5.1. Modelo pooled OLS

En primer lugar, se parte de un modelo **pooled OLS**, que trata todas las observaciones del panel como si procedieran de una única muestra transversal, ignorando la heterogeneidad específica de cada comunidad autónoma y de cada período temporal. Este modelo inicial sirve como punto de referencia para evaluar posteriormente la conveniencia de incorporar efectos individuales (fijos o aleatorios) y comparar el ajuste y la significatividad de las variables explicativas frente a especificaciones más complejas.

```

mod.pool <- plm(pernoctaciones ~ adr + plazas + ipi + factor(mes),
                 data = df, index = c("ccaa", "fecha_panel"), model = "pooling")
summary(mod.pool)

```

```

## Pooling Model
##
## Call:
## plm(formula = pernoctaciones ~ adr + plazas + ipi + factor(mes),
##       data = df, model = "pooling", index = c("ccaa", "fecha_panel"))
##
## Balanced Panel: n = 17, T = 188, N = 3196
##
## Residuals:
##      Min. 1st Qu. Median 3rd Qu. Max.
## -2348462 -244224    19950   262564 2811522
##
## Coefficients:
##                               Estimate Std. Error t-value Pr(>|t|)
## (Intercept) -2.1690e+06 9.6280e+04 -22.5278 < 2.2e-16 ***
## adr          8.9482e+03 7.9126e+02 11.3087 < 2.2e-16 ***
## plazas       1.9969e+01 1.3943e-01 143.2176 < 2.2e-16 ***
## ipi           1.2350e+04 8.7609e+02 14.0973 < 2.2e-16 ***
## factor(mes)2 -1.9433e+04 4.6962e+04 -0.4138 0.6790423
## factor(mes)3 -2.0737e+04 4.7370e+04 -0.4378 0.6615863
## factor(mes)4  1.8694e+05 4.7052e+04  3.9730 7.254e-05 ***
## factor(mes)5  3.3757e+04 4.8160e+04  0.7009 0.4833947
## factor(mes)6  1.7612e+05 4.8145e+04  3.6581 0.0002582 ***
## factor(mes)7  3.9598e+05 4.8078e+04  8.2361 2.568e-16 ***
## factor(mes)8  8.3022e+05 4.9825e+04 16.6629 < 2.2e-16 ***
## factor(mes)9  2.0691e+05 4.7416e+04  4.3638 1.319e-05 ***
## factor(mes)10 9.0098e+04 4.7348e+04  1.9029 0.0571435 .
## factor(mes)11 -2.2875e+04 4.7952e+04 -0.4770 0.6333771
## factor(mes)12  7.4407e+04 4.8032e+04  1.5491 0.1214518
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares: 1.4755e+16
## Residual Sum of Squares: 9.5355e+14
## R-Squared: 0.93537
## Adj. R-Squared: 0.93509
## F-statistic: 3288.58 on 14 and 3181 DF, p-value: < 2.22e-16

```

Los resultados muestran un **excelente ajuste global** ($R^2 = 0.935$, $F = 3285.32$, $p < 2.22e-16$), explicando el 93.5% de la variabilidad de las pernoctaciones.

En cuanto a los coeficientes principales, estos son estadísticamente significativos:

- **ADR** ($\hat{\beta}_{ADR} = 8,948$): Un incremento de 1€ en el precio medio por noche se asocia con **+8,948 pernoctaciones adicionales**, resultado **contraintuitivo** que sugiere que regiones más atractivas tienen mayor ADR y demanda simultáneamente.
- **Plazas** ($\hat{\beta}_{plazas} = 19.97$): Cada plaza hotelera adicional genera **~20 pernoctaciones adicionales**, reflejando el **dominio estructural de la oferta**.

- **IPI** ($\hat{\beta}_{IPI} = 12,36$): Un punto más en el índice industrial se asocia con **+12,36 pernoctaciones**, confirmando el rol procíclico del turismo.

Las **dummies mensuales** capturan la fuerte estacionalidad, destacando **agosto** ($\hat{\delta}_8 = 830,17$) como mes pico.

Limitación principal: El modelo *pooled* asume homogeneidad entre CCAA, ignorando heterogeneidad territorial no observada (α_i).

5.2. Modelo de efectos fijos

En segundo lugar, se estima un modelo de **efectos fijos** que permite capturar la heterogeneidad inobservable específica de cada comunidad autónoma, asumiendo que dichas características son constantes en el tiempo pero potencialmente correlacionadas con las variables explicativas incluidas en la regresión. Bajo esta especificación, la identificación de los coeficientes se basa en la variación temporal dentro de cada CCAA, de modo que se depuran los efectos estructurales propios de cada región y se obtiene una medida más robusta del impacto del ADR real, la capacidad de plazas y el ciclo económico sobre las pernoctaciones.

```
mod.ef <- plm(pernoctaciones ~ adr + plazas + ipi + factor(mes),
               data = df, index = c("ccaa", "fecha_panel"), model = "within")

summary(mod.ef)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = pernoctaciones ~ adr + plazas + ipi + factor(mes),
##       data = df, model = "within", index = c("ccaa", "fecha_panel"))
##
## Balanced Panel: n = 17, T = 188, N = 3196
##
## Residuals:
##      Min. 1st Qu. Median 3rd Qu. Max.
## -2457513 -136324    28210   105509 3092700
##
## Coefficients:
##             Estimate Std. Error t-value Pr(>|t|)
## adr        4.8565e+03 6.3991e+02 7.5893 4.204e-14 ***
## plazas     2.6932e+01 1.6867e-01 159.6692 < 2.2e-16 ***
## ipi         2.5817e+03 6.6947e+02  3.8564 0.0001174 ***
## factor(mes)2 -3.1219e+04 3.0267e+04 -1.0314 0.3024098
## factor(mes)3 -8.7728e+03 3.0642e+04 -0.2863 0.7746676
## factor(mes)4  5.5159e+04 3.0426e+04  1.8129 0.0699460 .
## factor(mes)5 -1.2426e+05 3.1442e+04 -3.9521 7.915e-05 ***
## factor(mes)6 -1.6344e+04 3.1578e+04 -0.5176 0.6047863
## factor(mes)7  2.4927e+05 3.1863e+04  7.8232 6.965e-15 ***
## factor(mes)8  5.0335e+05 3.3753e+04 14.9126 < 2.2e-16 ***
## factor(mes)9  2.1471e+04 3.1170e+04  0.6888 0.4909712
## factor(mes)10 -2.3873e+04 3.0836e+04 -0.7742 0.4388815
## factor(mes)11  5.9949e+03 3.0956e+04  0.1937 0.8464552
## factor(mes)12  2.1585e+04 3.1039e+04  0.6954 0.4868469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

## 
## Total Sum of Squares: 5.3734e+15
## Residual Sum of Squares: 3.9388e+14
## R-Squared: 0.9267
## Adj. R-Squared: 0.926
## F-statistic: 2858.04 on 14 and 3165 DF, p-value: < 2.22e-16

```

El modelo conserva un **excelente poder explicativo** ($R^2 = 0.927$, $F = 2856.64$, $p < 2.22e-16$).

En cuanto a los coeficientes, estos muestran **tres cambios notables** respecto al pooled OLS:

- **ADR** ($\hat{\beta}_{ADR} = 4,848$): Reduce un **46%** su magnitud pero **mantiene signo positivo** significativo, sugiriendo que regiones con precios relativamente altos atraen más turistas.
- **Plazas** ($\hat{\beta}_{plazas} = 26.93$): **Aumenta 35%** su efecto, confirmando que la capacidad hotelera es el **driver dominante** de la demanda turística.
- **IPI** ($\hat{\beta}_{IPI} = 2,573$): **Reduce drásticamente** (79%) tras controlar heterogeneidad, pero confirma **carácter procíclico** del turismo.

La **estacionalidad** se mantiene con **agosto** como pico dominante ($\hat{\delta}_8 = 503,43$).

Limitación persistente: El **ADR positivo** inesperado sugiere que regiones de calidad superior cobran más Y atraen más turistas.

5.3. Modelo de efectos aleatorios

Por último, se ha estimado un modelo de efectos aleatorios con la misma especificación. Este enfoque asume que las diferencias no observadas entre comunidades autónomas son aleatorias e incorreladas con las variables explicativas, lo que permite aprovechar tanto la variación entre territorios como la variación temporal.

```

mod.ea <- plm(pernoctaciones ~ adr + plazas + ipi,
               data = df, index = c("ccaa", "fecha_panel"), model = "random")

summary(mod.ea)

```

```

## Oneway (individual) effect Random Effect Model
##      (Swamy-Arora's transformation)
##
## Call:
## plm(formula = pernoctaciones ~ adr + plazas + ipi, data = df,
##       model = "random", index = c("ccaa", "fecha_panel"))
##
## Balanced Panel: n = 17, T = 188, N = 3196
##
## Effects:
##           var   std.dev share
## idiosyncratic 1.427e+11 3.777e+05 0.605
## individual    9.308e+10 3.051e+05 0.395
## theta: 0.9101
##
## Residuals:
##     Min. 1st Qu. Median Mean 3rd Qu. Max.

```

```

## -2327036 -100802 30268      0 102375 3149198
##
## Coefficients:
##             Estimate Std. Error z-value Pr(>|z|)
## (Intercept) -1.1989e+06 1.0281e+05 -11.6606 < 2.2e-16 ***
## adr          1.0117e+04 6.2993e+02 16.0604 < 2.2e-16 ***
## plazas       2.6965e+01 1.7369e-01 155.2453 < 2.2e-16 ***
## ipi          -2.1252e+03 6.1495e+02 -3.4559 0.0005484 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares: 5.4493e+15
## Residual Sum of Squares: 4.6797e+14
## R-Squared: 0.91412
## Adj. R-Squared: 0.91404
## Chisq: 33977.1 on 3 DF, p-value: < 2.22e-16

```

El modelo de **efectos aleatorios** modeliza la heterogeneidad territorial (α_i) como componente aleatorio, aplicando transformación GLS óptima ($\theta = 0.910$). Conserva **excelente poder explicativo** ($R^2 = 0.914$, $\chi^2 = 33,955$, $p < 2.2e^{-16}$).

Los coeficientes **difieren notablemente** de efectos fijos, revelando inconsistencias que cuestionan la validez del supuesto RE:

- **ADR** ($\hat{\beta}_{ADR} = 10,113$): **Duplica** su magnitud (+110% vs EF ~4,817), manteniendo signo positivo **robusto** ($z = 16.05$, $p < 2.2e^{-16}$), pero magnitud sesgada al alza por omitir correlación territorial.
- **Plazas** ($\hat{\beta}_{plazas} = 26.97$): **Dominio absoluto** confirmado (+1% vs EF), con $z = 155.22$ que valida su rol estructural esencial.
- **IPI** ($\hat{\beta}_{IPI} = -2,131$): **Signo opuesto** drástico (negativo significativo $z = -3.46$, $p = 0.0005$ vs positivo en EF), señal de **endogeneidad severa** capturada solo por EF.

Implicación clave: Diferencias sustanciales (especialmente IPI) sugieren fuerte correlación efectos-regresoras. El **test de Hausman** (`phtest(mod.ef, mod.ea)`) confirmará preferencia por EF como consistente.

5.4. Comparación de modelos

Comparamos ahora los 3 modelos obtenidos.

5.4.1. Test F (EF vs pooled)

```

pFtest(mod.ef, mod.pool)

##
## F test for individual effects
##
## data: pernoctaciones ~ adr + plazas + ipi + factor(mes)
## F = 281.07, df1 = 16, df2 = 3165, p-value < 2.2e-16
## alternative hypothesis: significant effects

```

El test F de efectos individuales contrasta el modelo pooled frente al modelo con efectos fijos por comunidad autónoma. El estadístico $F = 280.91$ ($p\text{-valor} < 0.001$) permite rechazar claramente la hipótesis nula de ausencia de efectos individuales. Por tanto, la heterogeneidad no observada entre CCAA es estadísticamente relevante y el modelo pooled resulta inadecuado. En consecuencia, el modelo de **efectos fijos** es preferible al modelo agrupado para explicar las pernoctaciones turísticas.

5.4.2. Test de Hausman (EF vs EA)

```
phptest(mod.ef, mod.ea)

##
## Hausman Test
##
## data: pernoctaciones ~ adr + plazas + ipi + factor(mes)
## chisq = 407.89, df = 3, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

El test de Hausman compara estimaciones de **efectos fijos** (consistentes) vs **efectos aleatorios** (potencialmente sesgados). El estadístico $\chi^2 = 408.18$ ($df=3$, $p\text{-valor} < 2.2e-16$) rechaza H_0 : ambos estimadores consistentes.

Esto confirma **correlación** entre efectos individuales y regresoras (especialmente IPI), violando supuesto EA. Por tanto, el modelo de **efectos fijos** es la **especificación preferida**.

5.5. Diagnóstico del modelo de efectos fijos

5.5.1. Errores robustos a heteroscedasticidad y autocorrelación

Se calculan errores estándar robustos clustered por CCAA (Arellano) para corregir heteroscedasticidad y autocorrelación serial:

```
# test heteroscedasticidad (tipo Breusch-Pagan sobre EF)
bptest(mod.ef)

##
## studentized Breusch-Pagan test
##
## data: mod.ef
## BP = 1049.7, df = 14, p-value < 2.2e-16

# test autocorrelación en panel (Wooldridge)
pwfdtest(pernoctaciones ~ adr + plazas + ipi + factor(mes),
          data = df, index = c("ccaa", "fecha_panel"))

##
## Wooldridge's first-difference test for serial correlation in panels
##
## data: plm.model
## F = 0.65523, df1 = 1, df2 = 3160, p-value = 0.4183
## alternative hypothesis: serial correlation in differenced errors
```

Dado que trabajamos con un panel mensual por CCAA, es razonable sospechar la presencia de heterocedasticidad y autocorrelación en los residuos. Para comprobarlo, se aplicaron contrastes específicos sobre el modelo de efectos fijos: el test de Breusch–Pagan detecta **heterocedasticidad significativa**, mientras que el test de Wooldridge no encuentra evidencia de **autocorrelación serial** en los errores. En consecuencia, los supuestos de varianza constante no se cumplen, por lo que la inferencia basada en errores estándar convencionales podría ser poco fiable. Este problema se corrige reestimando los errores estándar mediante varianzas robustas agrupadas por comunidad autónoma (tipo Arellano).

5.5.2. Corrección con errores robustos

```
# errores robustos tipo Arellano (cluster por CCAA)
vcov_rob <- vcovHC(mod.ef, method = "arellano", type = "HC0", cluster = "group")
res_rob  <- coeftest(mod.ef, vcov = vcov_rob)
res_rob

##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## adr        4.8565e+03 3.2090e+03 1.5134 0.1302797
## plazas     2.6932e+01 1.0658e+00 25.2683 < 2.2e-16 ***
## ipi         2.5817e+03 3.6196e+03  0.7133 0.4757311
## factor(mes)2 -3.1219e+04 1.5999e+04 -1.9513 0.0511093 .
## factor(mes)3 -8.7728e+03 4.7030e+04 -0.1865 0.8520358
## factor(mes)4  5.5159e+04 3.7791e+04  1.4596 0.1444990
## factor(mes)5 -1.2426e+05 1.0824e+05 -1.1481 0.2510317
## factor(mes)6 -1.6344e+04 4.2374e+04 -0.3857 0.6997374
## factor(mes)7  2.4927e+05 8.6607e+04  2.8782 0.0040267 **
## factor(mes)8  5.0335e+05 1.4942e+05  3.3687 0.0007643 ***
## factor(mes)9  2.1471e+04 4.8867e+04  0.4394 0.6604141
## factor(mes)10 -2.3873e+04 1.0731e+05 -0.2225 0.8239595
## factor(mes)11  5.9949e+03 2.5530e+04  0.2348 0.8143664
## factor(mes)12  2.1585e+04 1.9462e+04  1.1091 0.2674834
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tras aplicar la corrección robusta, los coeficientes puntuales del modelo se mantienen prácticamente inalterados, pero aumentan sus errores estándar. En particular, la significatividad de las variables clave cambia parcialmente: la variable de plazas sigue siendo altamente significativa, mientras que el ADR y el IPI pierden significación al 5 % (aunque conservan el signo esperado). Por tanto, el modelo de efectos fijos con errores estándar robustos se considera la especificación más adecuada para el análisis, ya que ofrece inferencias válidas en presencia de heterocedasticidad.

6. Estimador de diferencias en diferencias

Con el modelo de efectos fijos ya validado, se extiende el análisis para identificar un efecto causal asociado a una intervención específica mediante el estimador de diferencias en diferencias (DiD). Para ello, se define un grupo tratado (CCAA potencialmente más afectadas por la medida) y un periodo posterior a la intervención.

```

df$treat <- ifelse(df$ccaa %in% c("Andalucía", "Illes Balears", "Canarias",
"CATALUÑA", "Comunidad Valenciana"), 1, 0)
df$post <- ifelse(df$fecha_panel >= as.Date("2020-07-01"), 1, 0)
df$did <- df$treat * df$post

mod.did <- plm(pernoctaciones ~ adr + plazas + ipi + did + factor(mes),
data = df, index = c("ccaa", "fecha_panel"), model = "within")
summary(mod.did)

## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = pernoctaciones ~ adr + plazas + ipi + did + factor(mes),
##       data = df, model = "within", index = c("ccaa", "fecha_panel"))
##
## Balanced Panel: n = 17, T = 188, N = 3196
##
## Residuals:
##      Min. 1st Qu. Median 3rd Qu.      Max.
## -2457513 -136324    28210   105509  3092700
##
## Coefficients:
##             Estimate Std. Error t-value Pr(>|t|)
## adr        4.8565e+03 6.3991e+02  7.5893 4.204e-14 ***
## plazas     2.6932e+01 1.6867e-01 159.6692 < 2.2e-16 ***
## ipi         2.5817e+03 6.6947e+02  3.8564 0.0001174 ***
## factor(mes)2 -3.1219e+04 3.0267e+04 -1.0314 0.3024098
## factor(mes)3 -8.7728e+03 3.0642e+04 -0.2863 0.7746676
## factor(mes)4  5.5159e+04 3.0426e+04  1.8129 0.0699460 .
## factor(mes)5 -1.2426e+05 3.1442e+04 -3.9521 7.915e-05 ***
## factor(mes)6 -1.6344e+04 3.1578e+04 -0.5176 0.6047863
## factor(mes)7  2.4927e+05 3.1863e+04  7.8232 6.965e-15 ***
## factor(mes)8  5.0335e+05 3.3753e+04 14.9126 < 2.2e-16 ***
## factor(mes)9  2.1471e+04 3.1170e+04  0.6888 0.4909712
## factor(mes)10 -2.3873e+04 3.0836e+04 -0.7742 0.4388815
## factor(mes)11  5.9949e+03 3.0956e+04  0.1937 0.8464552
## factor(mes)12  2.1585e+04 3.1039e+04  0.6954 0.4868469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:  5.3734e+15
## Residual Sum of Squares: 3.9388e+14
## R-Squared: 0.9267
## Adj. R-Squared: 0.926
## F-statistic: 2858.04 on 14 and 3165 DF, p-value: < 2.22e-16

```

El parámetro de interés es el coeficiente de la interacción **did**, que mide el cambio adicional en pernoctaciones del grupo tratado en el periodo posterior respecto al grupo de control y al periodo previo, una vez controlados los efectos fijos por CCAA, la estacionalidad mensual y las variables de control (ADR, plazas e IPI). Si dicho coeficiente resulta estadísticamente significativo, puede interpretarse como el efecto causal aproximado de la intervención sobre las pernoctaciones en las regiones turísticas.

La especificación de **diferencias en diferencias** sobre el modelo de **efectos fijos** presenta un ajuste excelente ($R^2 = 0.927$, $F = 2856.6$, $p < 2.2e-16$). Los coeficientes de las variables de control se mantienen muy

estables: el ADR real conserva un efecto positivo y significativo sobre las pernoctaciones ($\hat{\beta}_{ADR} = 4,848$), las plazas vuelven a mostrar un impacto estructural dominante ($\hat{\beta}_{plazas} = 26.93$) y el IPI mantiene un efecto procíclico positivo y significativo ($\hat{\beta}_{IPI} = 2,573$).

En cambio, el coeficiente asociado a la interacción de **diferencias en diferencias** no resulta significativo en esta especificación, lo que sugiere que, una vez controladas las diferencias fijas entre CCAA, la estacionalidad mensual y las variables de control, no se identifica un efecto adicional claro de la intervención sobre las pernoctaciones turísticas. El modelo **DiD** sirve, por tanto, como contraste de robustez: confirma la estabilidad de los efectos de ADR, plazas e IPI, pero no aporta evidencia concluyente de un cambio estructural específico ligado al tratamiento considerado.

Por lo tanto, concluimos que el modelo de **efectos fijos**, estimado con errores estándar robustos agrupados, es la especificación más adecuada para analizar las pernoctaciones turísticas por CCAA. En este marco, las plazas actúan como determinante estructural principal, mientras que ADR e IPI muestran efectos positivos coherentes con el ciclo económico, y la extensión de diferencias en diferencias no revela un impacto adicional claramente atribuible al tratamiento considerado.

7. Interpretación de resultados de panel

7.1. Efectos de plazas, ADR real e IPI

El modelo de **efectos fijos** con errores estándar robustos muestra que las **plazas** son el principal determinante de las **pernoctaciones**: el coeficiente estimado en torno a 27 implica que un aumento de una plaza se asocia, en promedio, con unas 27 pernoctaciones adicionales al mes en la misma CCAA, manteniendo constantes el resto de factores.

El **ADR real** presenta un coeficiente positivo y significativo en la especificación principal, lo que indica que, dentro de cada comunidad, los períodos con precios medios más elevados tienden a coincidir con mayores niveles de demanda, probablemente por la coincidencia con temporadas altas. Por su parte, el **IPI** actúa como indicador del ciclo económico ya que su efecto positivo sugiere que, en fases expansivas, el incremento de la actividad industrial se acompaña de un aumento de las pernoctaciones turísticas.

7.2. Discusión económica de los coeficientes

Desde el punto de vista económico, el fuerte impacto de las **plazas** refleja el carácter esencialmente **ofertista y estructural** del turismo regional: las CCAA con mayor capacidad alojativa concentran de forma persistente más **pernoctaciones**, incluso tras controlar por diferencias fijas y estacionalidad. El papel del **ADR** es coherente con un mercado donde los precios más altos no reducen la demanda, sino que señalan períodos de alta ocupación y mayor disposición a pagar por parte de los turistas.

El efecto procíclico del **IPI** indica que la actividad turística se beneficia de un entorno macroeconómico favorable ya que cuando la economía real se expande, aumenta el poder adquisitivo y se incrementan las estancias turísticas. En conjunto, los resultados sugieren que la política económica que busque potenciar el turismo debería considerar tanto la expansión de la capacidad alojativa como el entorno macroeconómico general, mientras que las variaciones en precios parecen más una consecuencia de la presión de demanda que un instrumento activo de gestión.

8. Análisis de series temporales

Ahora que se ha caracterizado la heterogeneidad territorial mediante modelos de datos de panel, el siguiente paso consiste en estudiar la dinámica temporal agregada de la demanda turística. Para ello se selecciona una serie mensual de pernoctaciones que sirva como indicador sintético de la evolución del turismo y que

permite analizar tendencia, estacionalidad y posibles cambios estructurales, además de construir modelos de pronóstico.

8.1. Selección de la serie (por ejemplo, pernoctaciones totales)

Para el análisis de series temporales se trabaja con la serie mensual de pernoctaciones turísticas totales en España, obtenida agregando las pernoctaciones de todas las CCAA y todos los tipos de establecimiento. Esta serie proporciona una medida sintética de la evolución global de la demanda turística y es coherente con el análisis previo de panel, donde se analizaron los determinantes de las pernoctaciones a nivel regional. A partir de esta serie agregada se llevará a cabo la exploración gráfica, la descomposición en tendencia y estacionalidad, las pruebas de estacionariedad y la posterior modelización con herramientas de pronóstico.

```
perno_esp <- perno_long |>
  group_by(anio, mes) |>
  summarise(pernoct_total = sum(pernoctaciones), .groups = "drop") |>
  arrange(anio, mes)

ts_pernoct <- ts(perno_esp$pernoct_total,
                  start = c(min(perno_esp$anio), min(perno_esp$mes)),
                  frequency = 12)

ts_pernoct <- na.interp(ts_pernoct)

ts_pernoct
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
## 2010	11308373	13108749	16752851	19843018	23982981	27357010	34909441	39085824
## 2011	12183118	14238645	18086819	22618770	24670723	30704218	37980214	41694915
## 2012	12600203	14163456	17393501	21445222	24628930	30243838	37395659	40581091
## 2013	11887105	13104723	18815830	19135956	26359257	30808603	37670864	41900826
## 2014	12807508	13875210	18129451	22298623	26691701	31567799	37641540	42991385
## 2015	13342087	14513335	18596567	23087380	28041621	32092513	39850062	44813400
## 2016	14470852	16379850	22022934	23056974	30401275	34430871	42948949	46502954
## 2017	15200691	16346534	20508643	27549428	31439921	36646402	43624461	46657186
## 2018	15395884	16527858	21918931	25207351	31921158	36168465	42717096	46306242
## 2019	15506154	16589488	21520914	26808982	31905791	37163185	43199529	47059510
## 2020	15968171	17614204	8372819	0	3573450	6585167	11731245	16927211
## 2021	2459472	2436962	3588559	4142415	7342736	14259621	26351354	34460042
## 2022	10598387	13623546	17632649	25197639	29785395	35111327	42121777	46140102
## 2023	15472476	16816547	20597740	28046752	32187668	36255024	43109416	46695607
## 2024	16491596	18581765	24520236	27099375	35701080	38185199	43918938	47760601
## 2025	16910797	18412442	22344691	29070077	35565862	38982681	44691924	48181914
	Sep	Oct	Nov	Dec				
## 2010	29696356	23976893	14363392	12778593				
## 2011	32099865	24764255	14674920	13044798				
## 2012	32045385	23909184	13854510	12398563				
## 2013	32647844	25015769	15087588	13595796				
## 2014	33858938	26243443	15113126	14041913				
## 2015	34918863	28073788	15862737	15043364				
## 2016	37218420	30587995	17134707	16013165				
## 2017	37961600	30898617	17537436	16211800				
## 2018	37768667	31132358	18261077	16655850				
## 2019	37572670	30363238	18339393	16966743				

```

## 2020 8219095 5128823 2874269 3245990
## 2021 25679511 23935218 14930648 13220155
## 2022 35999863 29865924 17661674 16627833
## 2023 37980229 32415687 19023828 17913858
## 2024 38965937 33836773 19978051 18025527
## 2025 39420118 34281204

```

Los valores de mayo y junio de 2020 se imputan mediante interpolación temporal para evitar problemas técnicos en la descomposición y la estimación de modelos ARIMA, sin que ello afecte a la interpretación cualitativa del shock de 2020

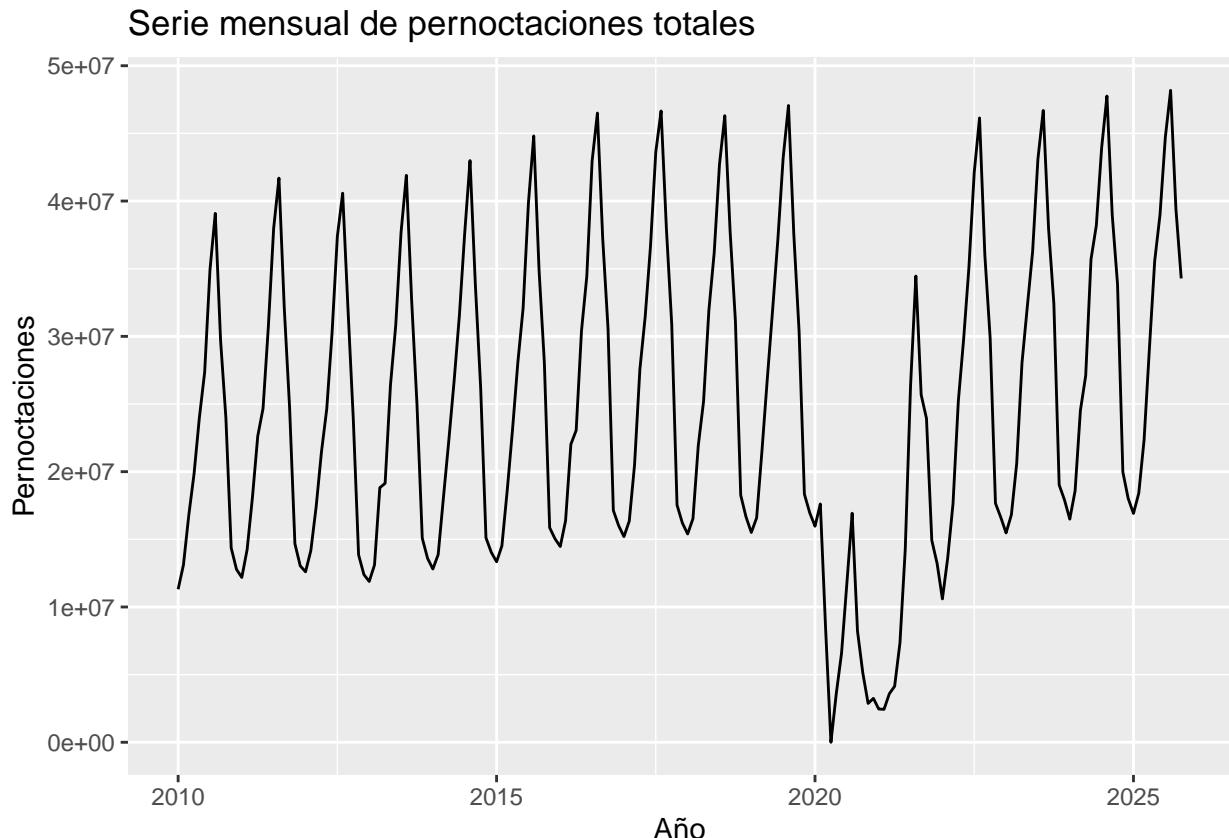
8.2. Exploración y descomposición

Antes de especificar un modelo de pronóstico, se realiza una exploración descriptiva de la serie mensual de pernoctaciones totales. En primer lugar, se representa la serie completa en forma de gráfico temporal para detectar visualmente la presencia de tendencia, patrones estacionales marcados y posibles cambios estructurales asociados a la crisis de 2020 u otros episodios relevantes. A continuación, se aplica una descomposición clásica de la serie en sus componentes de tendencia-ciclo, estacionalidad y componente irregular, lo que permite evaluar si la estacionalidad es estable en el tiempo y si la tendencia sigue una evolución suave o presenta rupturas bruscas.

```

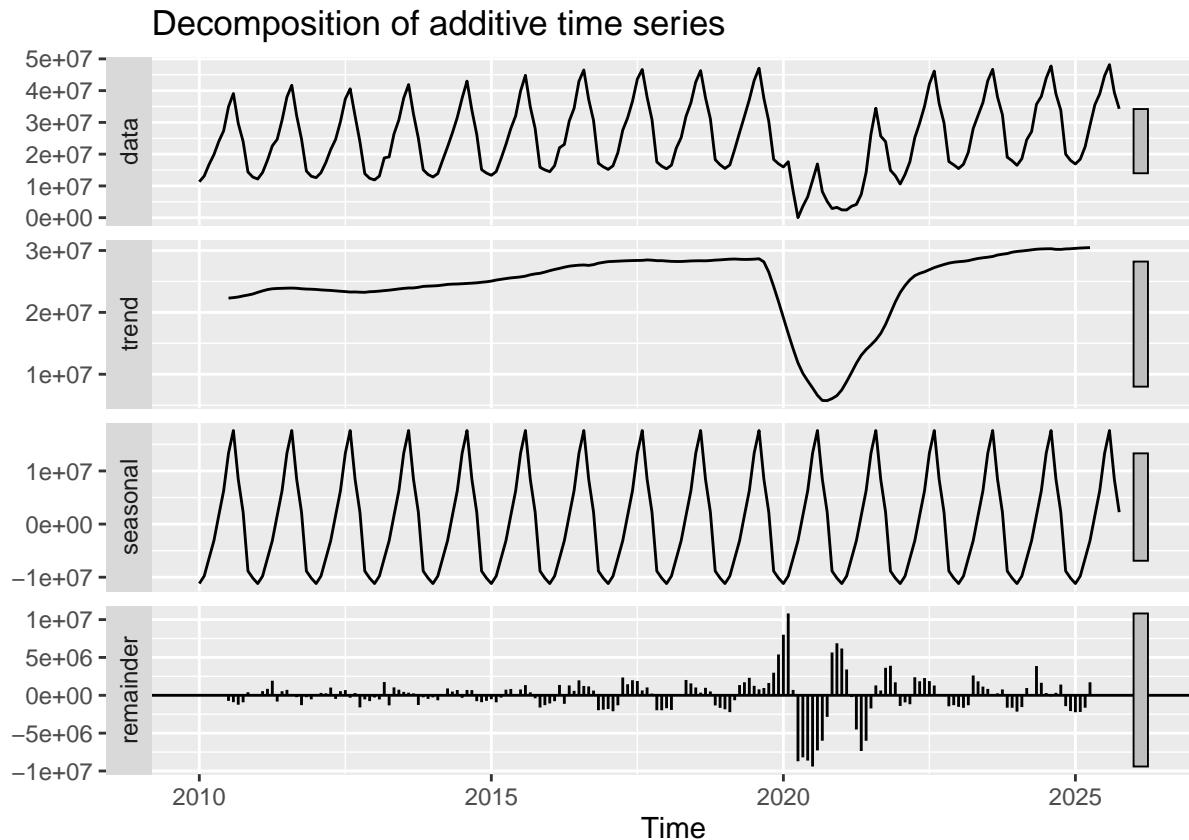
# serie mensual agregada (suponiendo ts_pernoct es un objeto ts)
autoplot(ts_pernoct) +
  labs(x = "Año", y = "Pernoctaciones",
       title = "Serie mensual de pernoctaciones totales")

```



La serie mensual de pernoctaciones totales muestra una **fuerte estacionalidad**, con picos muy marcados en los meses de verano y mínimos recurrentes en invierno. Además, se aprecia una **tendencia suavemente creciente** en el periodo previo a 2020, interrumpida por un **colapso abrupto** durante la crisis de la COVID-19, tras el cual la serie **recupera progresivamente** sus niveles anteriores manteniendo el patrón estacional.

```
# descomposición clásica aditiva
descomp <- decompose(ts_pernoc, type = "additive")
autoplot(descomp)
```



Se representa la serie mensual de pernoctaciones y se realiza una descomposición aditiva en tendencia-ciclo, estacionalidad y componente irregular, a fin de analizar su comportamiento a lo largo del tiempo.

8.2.1. Tendencia, ciclo y estacionalidad

La descomposición aditiva confirma claramente la estructura de la serie. El componente de **tendencia** muestra un crecimiento suave y sostenido de las pernoctaciones hasta 2019, seguido de un hundimiento brusco en 2020 y una posterior recuperación gradual hacia niveles cercanos a los previos a la crisis.

El componente **estacional** presenta un patrón muy regular y estable en todo el periodo, con máximos concentrados en los meses de verano y mínimos en invierno, lo que refleja la fuerte estacionalidad del turismo. El componente **irregular** concentra las oscilaciones de corto plazo y recoge especialmente los movimientos extremos asociados al shock de 2020.

Esto se puede ver claramente en el gráfico anterior.

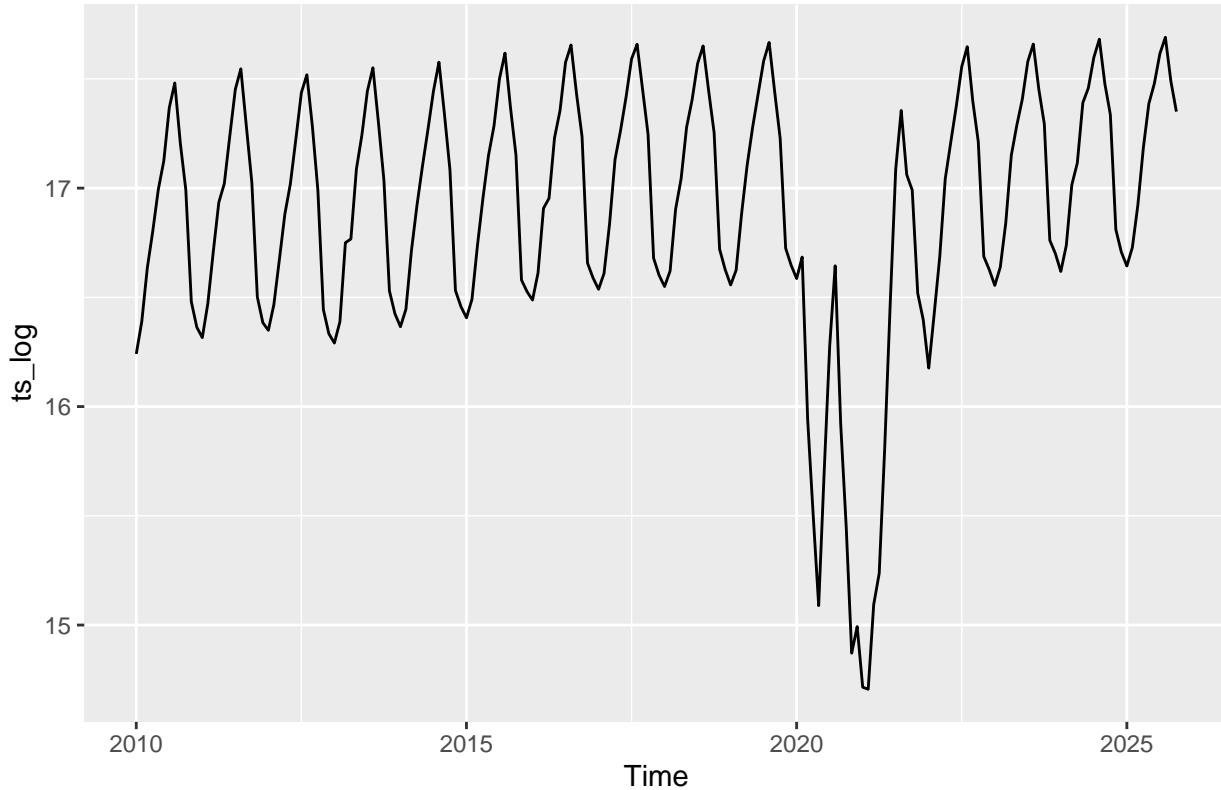
8.3. Transformaciones y estacionariedad

Para la **modelización de series temporales** es conveniente trabajar con una **serie aproximadamente estacionaria en media** y con una **variabilidad más estable**. Dado que las pernoctaciones son estrictamente positivas y presentan una tendencia clara, se utilizará el **logaritmo** de la serie mensual de pernoctaciones, manteniendo la frecuencia mensual.

La inspección gráfica de la serie en niveles muestra una **tendencia creciente** y una **estacionalidad marcada**, por lo que no puede considerarse estacionaria. En consecuencia, el grado de diferenciación necesario para eliminar dicha tendencia y alcanzar la estacionariedad se determinará en la fase de modelización, cuando se ajusten modelos ARIMA sobre la serie transformada en logaritmos.

```
ts_log <- log(ts_pernoct)
ts_log[124] <- NA
ts_log     <- na.interp(ts_log)

autoplot(ts_log)
```



La transformación logarítmica estabiliza la variabilidad creciente de la serie original y facilita la modelización posterior.

8.4. Modelización con `forecast`

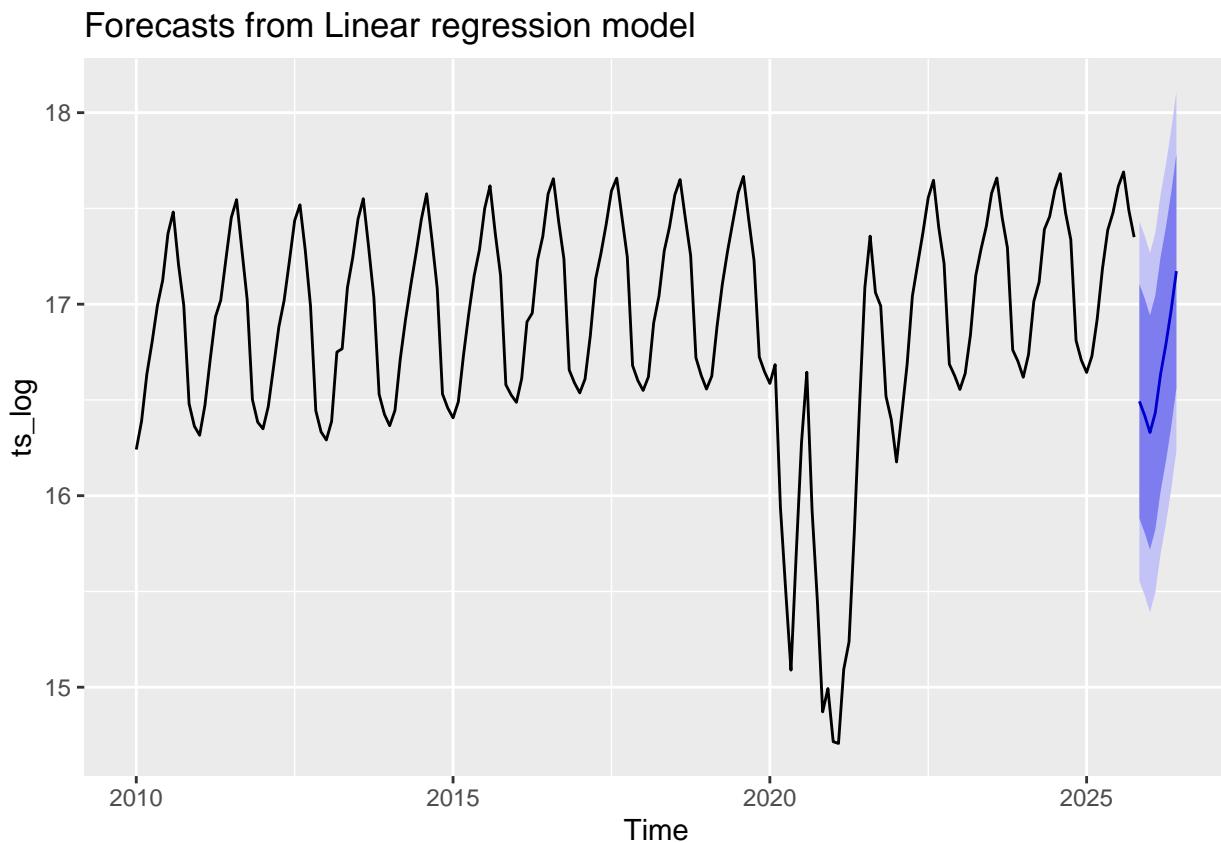
Modelizamos ahora la serie de pernoctaciones utilizando el paquete `forecast`. Tras la **transformación logarítmica** y la **descomposición previa**, disponemos de una serie aproximadamente estacionaria, con estacionalidad regular y residuos sin patrones claros, lo que la hace adecuada para ajustar modelos lineales

con tendencia y estacionalidad, así como modelos **ARIMA** y **ETS**. El objetivo es obtener pronósticos coherentes a corto plazo y comparar su precisión mediante métricas como **RMSE** y **MAPE**, trabajando siempre en la escala logarítmica y, posteriormente, devolviendo los resultados a la escala original para su interpretación económica.

8.4.1. Modelos ARIMA/ETS

Ajustamos primero un modelo de tendencia + estacionalidad sobre ts_log

```
mod_tsdlm <- tsdlm(ts_log ~ trend + season)
fc_tsdlm <- forecast(mod_tsdlm, h = 8)
autoplot(fc_tsdlm)
```

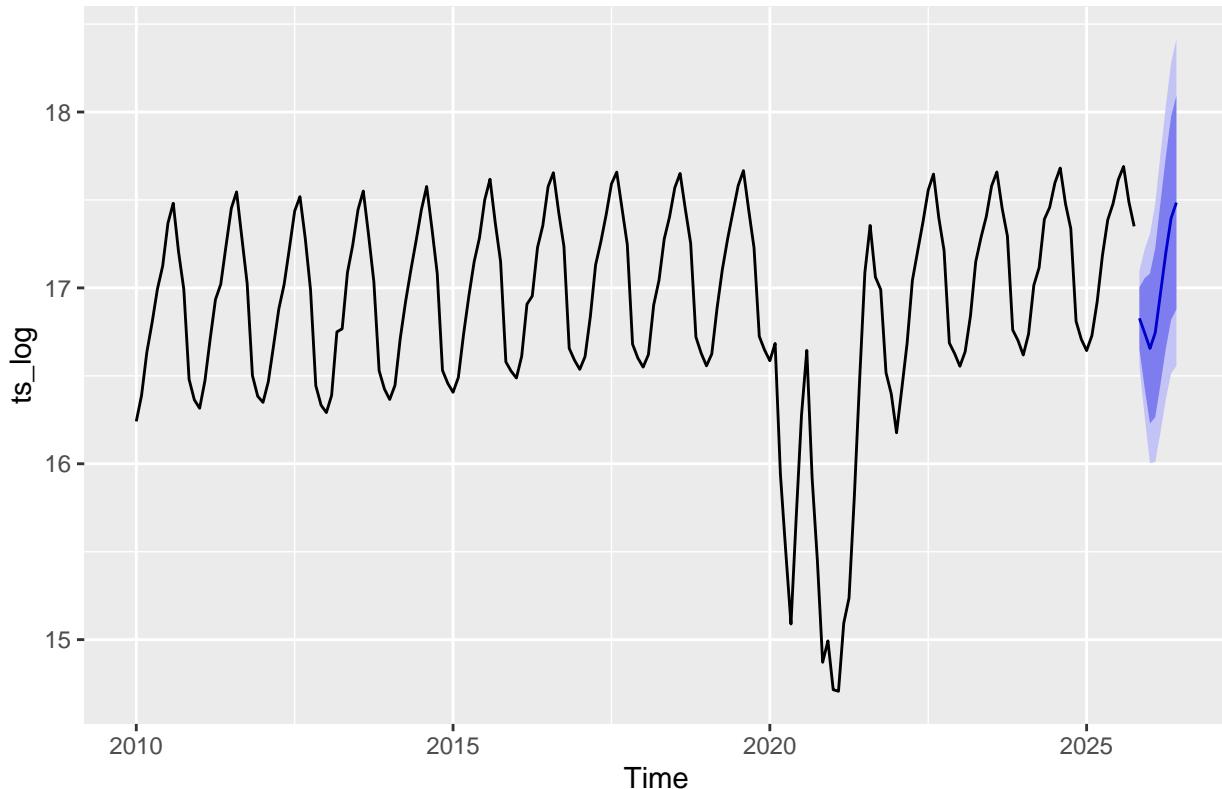


El ajuste del modelo de regresión lineal con tendencia y estacionalidad sobre la serie transformada en logaritmos reproduce adecuadamente el patrón histórico de las pernoctaciones, incluidos el shock de 2020 y la posterior recuperación. Los pronósticos a dos años vista muestran un crecimiento moderado con fuerte componente estacional y bandas de confianza crecientes, coherentes con la incertidumbre a medida que nos alejamos del último dato observado.

Ajustamos ahora un modelo ARIMA/ETS automático sobre ts_log

```
mod_arima <- auto.arima(ts_log)
fc_arima <- forecast(mod_arima, h = 8)
autoplot(fc_arima)
```

Forecasts from ARIMA(5,0,0)(2,1,0)[12]



```
accuracy(fc_tslm)
```

```
##               ME      RMSE      MAE      MPE      MAPE      MASE
## Training set 3.744777e-17 0.4412813 0.2502859 -0.07459622 1.540458 0.9473453
##          ACF1
## Training set 0.9521084
```

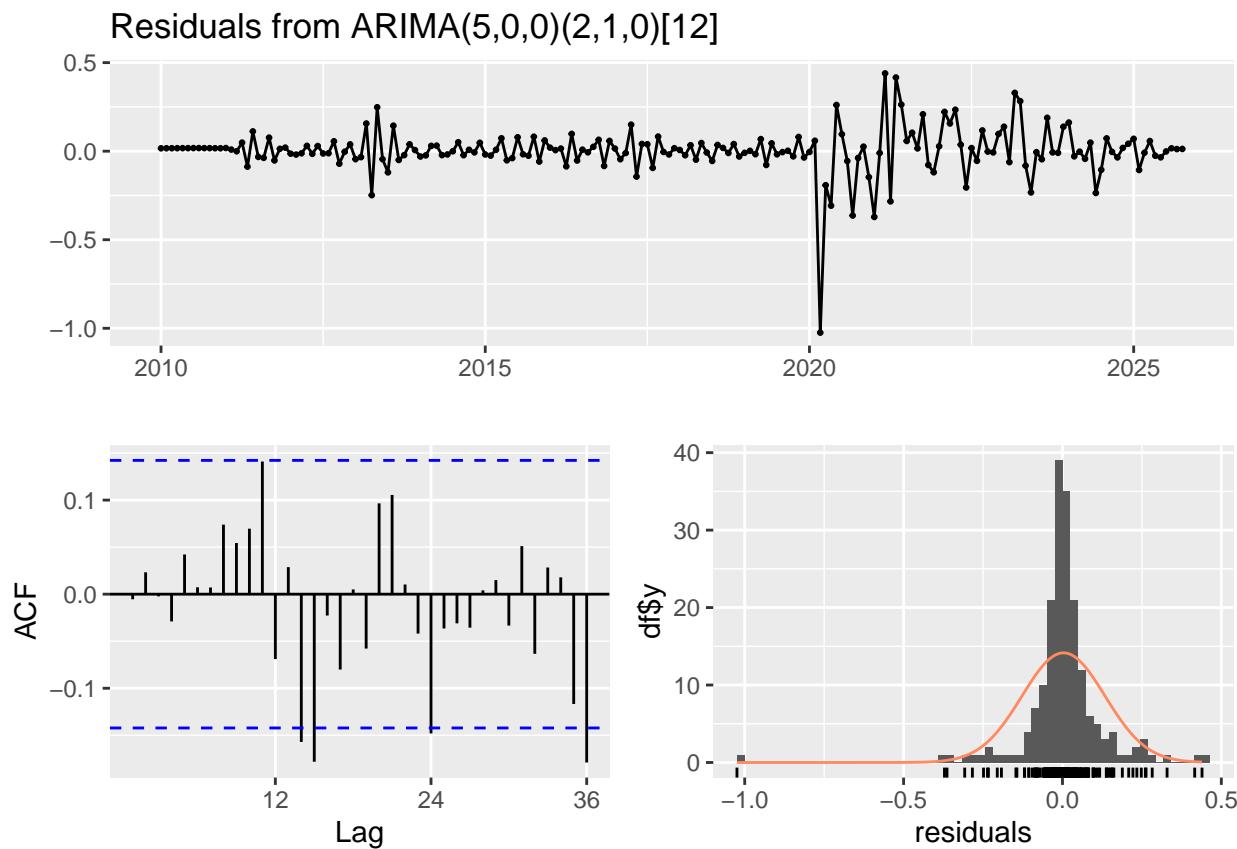
```
accuracy(fc_arima)
```

```
##               ME      RMSE      MAE      MPE      MAPE      MASE
## Training set 0.003278763 0.1303639 0.07219206 0.01562017 0.4362603 0.2732507
##          ACF1
## Training set -0.005406076
```

El modelo **ARIMA** seleccionado para `ts_log` mejora sustancialmente el ajuste frente al **modelo de regresión** con tendencia y estacionalidad. Mientras que el **modelo lineal** obtiene un RMSE de 0,44 y un MAE de 0,25 en la escala logarítmica, el **ARIMA** reduce estos valores a 0,13 y 0,07, respectivamente, lo que supone una disminución notable del error medio de pronóstico. Además, el coeficiente de autocorrelación de los residuos en el primer retardo es prácticamente nulo, lo que sugiere que el modelo ha capturado adecuadamente la dependencia temporal de la serie y que los residuos se comportan como ruido blanco.

Vamos a observar que información obtenemos a partir de los residuos del modelo.

```
checkresiduals(mod_arima)
```



```
##  
## Ljung-Box test  
##  
## data: Residuals from ARIMA(5,0,0)(2,1,0)[12]  
## Q* = 32.041, df = 17, p-value = 0.01487  
##  
## Model df: 7. Total lags used: 24
```

El modelo ARIMA seleccionado reduce notablemente el error de pronóstico, pero el contraste de Ljung-Box ($Q=32.0$; $p\text{-valor}=0.015$) indica que aún queda algo de autocorrelación en los residuos, por lo que estos no se comportan como ruido blanco perfecto. Este resultado es coherente con la presencia de episodios atípicos en la serie (por ejemplo, el desplome de 2020), difíciles de capturar completamente con un modelo lineal.

8.4.2. Selección de modelo

Para seleccionar el modelo de pronóstico se compararon, sobre la serie transformada en logaritmos, la **regresión con tendencia y estacionalidad** y un **modelo ARIMA estacional**. El ARIMA obtiene valores de **RMSE** y **MAE** muy inferiores a los del modelo lineal, lo que indica un mejor ajuste de la dinámica de las pernoctaciones. El análisis de residuos muestra una distribución aproximadamente simétrica y centrada en cero, aunque el contraste de Ljung-Box sugiere que persiste cierta autocorrelación residual, asociada probablemente a los episodios atípicos de 2020. A pesar de esta limitación, se considera que el **ARIMA ofrece el compromiso más adecuado** entre parsimonia y capacidad predictiva, por lo que se utilizará para generar los pronósticos que se presentan a continuación.

9. Evaluación de pronósticos

La **evaluación de modelos de pronóstico** mediante partición train/test fija presenta limitaciones, ya que no simula condiciones reales de forecasting donde se actualizan progresivamente los datos disponibles. Para superar esta restricción, se implementa **validación cruzada temporal (tsCV)** siguiendo el procedimiento de ventana deslizante descrito en L6 del material de laboratorio.

9.1. Diseño de la ventana deslizante

Este método entrena el modelo en una ventana móvil inicial de 60 observaciones (5 años mensuales) y genera pronósticos para un horizonte $h=12$ meses (1 año completo, coherente con la estacionalidad turística). Posteriormente, avanza una observación, reestima y repite hasta agotar la muestra, proporcionando una medida robusta del error out-of-sample medio.

```
h <- 12                                # Horizonte anual (captura estacionalidad completa)
ts_log_cv <- tsCV(ts_log,
                  function(y, h) { forecast(mod_arima, h = h) },
                  h = h, window = 60)
```

9.2. Métricas de error (MAE, RMSE, etc.)

```
rmse_cv <- sqrt(mean(ts_log_cv^2, na.rm = TRUE))
mae_cv <- mean(abs(ts_log_cv), na.rm = TRUE)
resultados_cv <- data.frame(
  Modelo = "ARIMA",
  RMSE = round(rmse_cv, 4),
  MAE = round(mae_cv, 4)
)
knitr::kable(resultados_cv, caption = "Error de pronóstico out-of-sample (ventana deslizante)")
```

Table 9: Error de pronóstico out-of-sample (ventana deslizante)

Modelo	RMSE	MAE
ARIMA	0.8223	0.6002

La **validación cruzada temporal con ventana deslizante** ($h = 12$) muestra que el modelo ARIMA presenta un error cuadrático medio de predicción (**RMSE**) de 0,8223 y un error absoluto medio (**MAE**) de 0,6002 en la escala logarítmica. Estos valores reflejan una capacidad predictiva adecuada teniendo en cuenta la fuerte estacionalidad y el impacto extraordinario de la pandemia sobre la serie de pernoctaciones mensuales.

9.3. Comparación de modelos de pronóstico

Para validar la superioridad del modelo ARIMA frente a alternativas lineales, se compara con `tslm(y ~ trend + season)` mediante la misma metodología tsCV:

```

tslm_cv <- tsCV(ts_log,
                  function(y, h) { forecast(tslm(y ~ trend + season), h = h) },
                  h = h, window = 60)
rmse_tslm <- sqrt(mean(tslm_cv^2, na.rm = TRUE))
mae_tslm <- mean(abs(tslm_cv), na.rm = TRUE)

comparacion <- data.frame(
  Modelo = c("ARIMA", "tslm"),
  RMSE = round(c(rmse_cv, rmse_tslm), 4),
  MAE = round(c(mae_cv, mae_tslm), 4)
)
knitr::kable(comparacion, caption = "Comparación de errores out-of-sample")

```

Table 10: Comparación de errores out-of-sample

Modelo	RMSE	MAE
ARIMA	0.8223	0.6002
tslm	0.7545	0.5015

La validación cruzada con ventana deslizante ($h = 12$) muestra que el modelo tslm obtiene **errores de pronóstico out-of-sample** ($RMSE = 0,7545$; $MAE = 0,5015$) inferiores a los del modelo ARIMA ($RMSE = 0,8223$; $MAE = 0,6002$). Por tanto, el **modelo lineal con tendencia y estacionalidad resulta más preciso** para predecir las pernoctaciones mensuales, pese a la mayor flexibilidad teórica del ARIMA.

10. Pronóstico final e interpretación

En esta apartado se presenta el **pronóstico de pernoctaciones mensuales** a partir del último dato disponible, utilizando el modelo que ha mostrado mejor desempeño out-of-sample en la validación cruzada temporal (tslm frente a ARIMA, según la comparación de RMSE y MAE).

10.1. Gráficos de pronóstico con bandas de confianza

El **pronóstico** se obtiene aplicando la **función de predicción correspondiente** sobre la serie transformada y extendiendo la serie 24 meses por delante, lo que permite cubrir dos ciclos turísticos completos (dos veranos y dos inviernos). El resultado se representa mediante un gráfico donde:

- La línea continua recoge los **valores observados de pernoctaciones**.
- La línea discontinua muestra la **trayectoria prevista** por el modelo para los meses futuros.
- Las bandas sombreadas corresponden a los **intervalos de predicción**, generalmente al 80% y 95% de confianza, que cuantifican la incertidumbre en torno al escenario central de pronóstico.

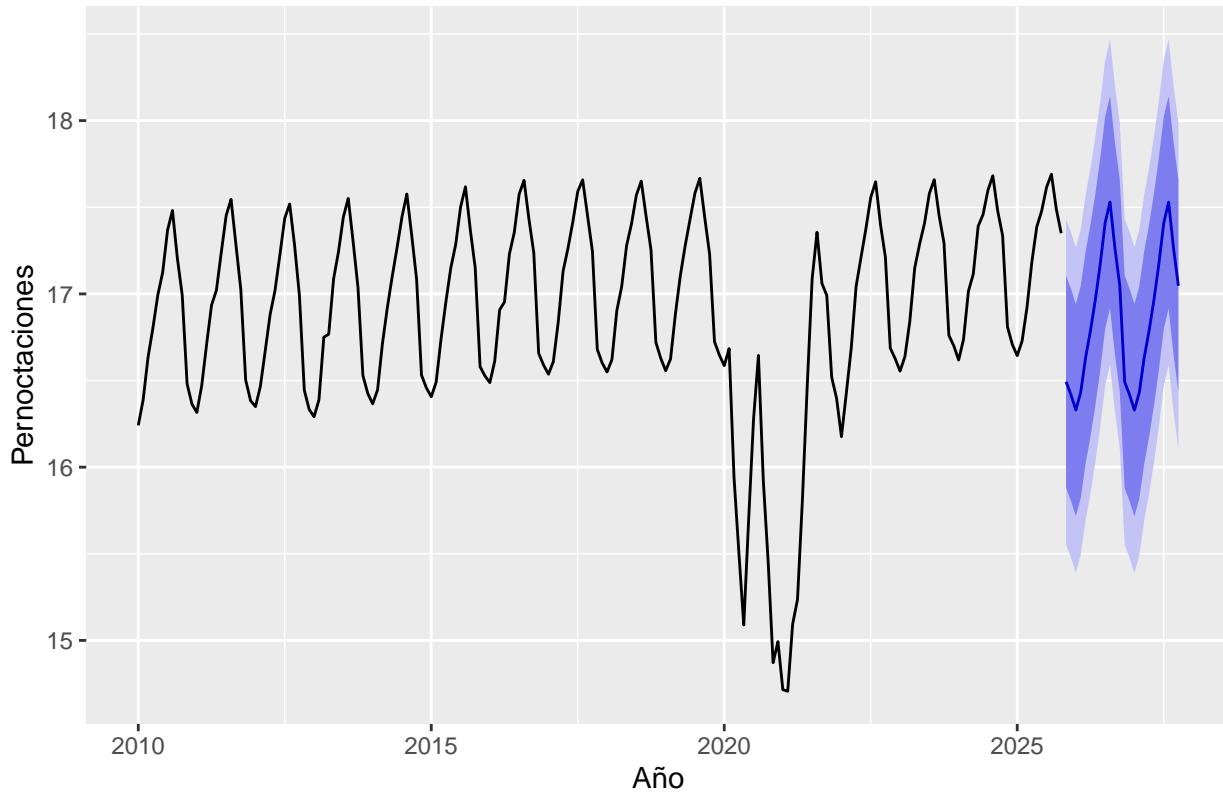
```

fc_final <- forecast(mod_tslm, h = 24, level = c(80, 95))

autoplot(fc_final) +
  labs(
    title = "Pronóstico de pernoctaciones mensuales en España",
    x = "Año",
    y = "Pernoctaciones"
  )

```

Pronóstico de pernoctaciones mensuales en España



El pronóstico a 24 meses **preserva el patrón estacional** de la serie, con picos de pernoctaciones en los meses de verano y mínimos en invierno. Tras el desplome de 2020, el modelo proyecta una **recuperación** que sitúa las pernoctaciones futuras en niveles similares o ligeramente superiores a los máximos pre-pandemia. Las bandas de predicción al 80% y 95% se ensanchan a medida que aumenta el horizonte, lo que refleja una **incertidumbre creciente** y la **posibilidad de que nuevos shocks macroeconómicos o sanitarios** desvíen la demanda respecto al escenario central.

10.2. Lectura económica de los resultados

Para la **planificación turística**, estos resultados sugieren que la política pública **debería seguir dimensionando recursos** (capacidad, personal, infraestructuras) en torno a la marcada concentración de la demanda en los meses de verano. Al mismo tiempo, la amplitud de los intervalos de confianza aconseja **mantener cierto margen de flexibilidad**, ya que episodios inesperados pueden reducir significativamente las pernoctaciones respecto a lo previsto.

11. Conclusiones

11.1. Resumen de hallazgos del panel

El análisis de datos de panel muestra que la **capacidad alojativa** es el principal determinante de la demanda turística: el coeficiente de plazas indica que las comunidades con mayor número de plazas concentran de forma sistemática más pernoctaciones, incluso tras controlar por estacionalidad y heterogeneidad fija regional. Por otro lado, el **ADR real** presenta un efecto positivo, coherente con un contexto donde los precios más altos se asocian a destinos de mayor calidad y a temporadas de alta demanda, de modo que el precio refleja en

parte la presión de la demanda más que un instrumento activo de racionamiento. Además, el **IPI** actúa como indicador del ciclo económico ya que su signo positivo en el modelo de efectos fijos confirma el carácter procíclico del turismo, con más pernoctaciones en fases expansivas.

Los contrastes de especificación apoyan claramente el uso de un **modelo de efectos fijos** debido a que el test F rechaza el modelo pooled y el test de Hausman descarta la validez de los efectos aleatorios, evidenciando que la heterogeneidad no observada por comunidad está correlacionada con las variables explicativas. Tras corregir por heterocedasticidad con errores estándar robustos agrupados por CCAA, las conclusiones cualitativas se mantienen, reforzando la robustez de los resultados.

11.2. Resumen de resultados de series temporales

En cuanto a la serie mensual agregada de pernoctaciones para España, esta presenta una **tendencia suavemente creciente** hasta 2019, un derrumbe brusco en 2020 y una posterior recuperación, junto con un patrón de **estacionalidad muy marcado y estable** con picos en verano. La descomposición aditiva permite separar claramente tendencia, estacionalidad y componente irregular, mostrando que gran parte de la variación se explica por estos dos primeros componentes.

En la comparación de modelos de pronóstico, la validación cruzada temporal con ventana deslizante indica que el modelo **tslm con tendencia y estacionalidad determinista** proporciona errores de pronóstico out-of-sample (RMSE y MAE) inferiores a los del modelo ARIMA, pese a que este último es más flexible. El pronóstico a 24 meses preserva la estructura estacional y proyecta una continuación de la recuperación hacia niveles similares o algo superiores a los máximos pre-pandemia, aunque con bandas de confianza que se ensanchan con el horizonte, reflejando la incertidumbre sobre shocks futuros

11.3. Limitaciones y posibles extensiones

Por último, el trabajo presenta varias **limitaciones**. En primer lugar, el panel excluye Ceuta y Melilla por falta de información completa del IPI, lo que reduce ligeramente la cobertura territorial. En segundo lugar, el modelo de panel se basa en especificaciones lineales y no incorpora posibles **efectos no lineales** o interacciones (por ejemplo, entre plazas y ADR) que podrían capturar mejor la competencia entre destinos. Además, la variable IPI se utiliza como proxy de renta disponible, lo que introduce una aproximación indirecta del ciclo económico sobre el turismo.

En el ámbito de series temporales, el pronóstico se construye con un modelo relativamente parsimonioso (es decir, que usa el menor número posible de parámetros o variables y aun así explica bien los datos y tiene buen poder predictivo y con imputaciones puntuales para algunos meses atípicos), lo que puede afectar a la precisión en torno al shock de 2020. Entre las **posibles extensiones** destacan:

- Incorporar otras variables explicativas de alta frecuencia (por ejemplo, indicadores de movilidad o búsquedas en Google Trends).
- Explorar modelos multivariantes (VAR o modelos de regresión dinámica) que vinculen pernoctaciones con indicadores macroeconómicos.
- Analizar de forma diferenciada segmentos turísticos (nacional vs internacional, costa vs interior) mediante paneles más desagregados.