ÍNDICE DE CONTENIDOS

1 Descripción general del proyecto

Este estudio trata de analizar el impacto que el flujo turístico tiene sobre el precio del suelo urbano en las provincias de España. Para ello, se llevará a cabo la integración y el análisis de dos bases de datos clave: una sobre los registros del número de viajeros y pernoctaciones por provincias españolas, y otra que contiene los valores medios del metro cuadrado de suelo urbano de cada provincia [1] [2].

No es un secreto que España es uno de los principales destinos turísticos del mundo, atrayendo aproximadamente 71,7 millones de turistas internacionales en 2022 y 85,1 millones en 2023 como se puede observar en la Figura 1 [3].

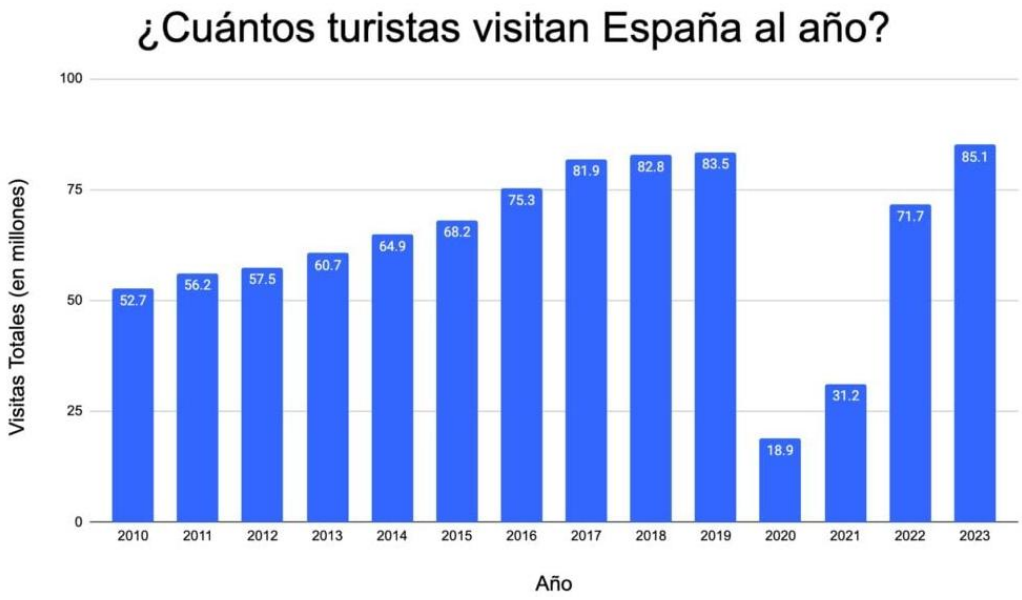

Figura 1. Gráfica de turistas que visitan España cada año. Fuente: (Road Genius, 2023)

Figura 2. Gráfica de turismo internacional vs nacional por visitas. Fuente: (Road Genius, 2023)

Los datos públicos más recientes son que llegaron al país 5,7 millones de turistas internacionales en noviembre, un 10,3% más que en el mismo mes de 2023. De manera que, según los datos del Instituto Nacional de Estadística (INE), los once primeros meses de 2024, el número de turistas internacionales que llegaron a España alcanzó su cifra más alta, superando los 88,5 millones. (Siendo Canarias (25,6% del total), Cataluña (22,2%) y Andalucía (13,7%) los principales destinos turísticos.). Estos datos se pueden observar en la Figura 3 sacada del INE [4].

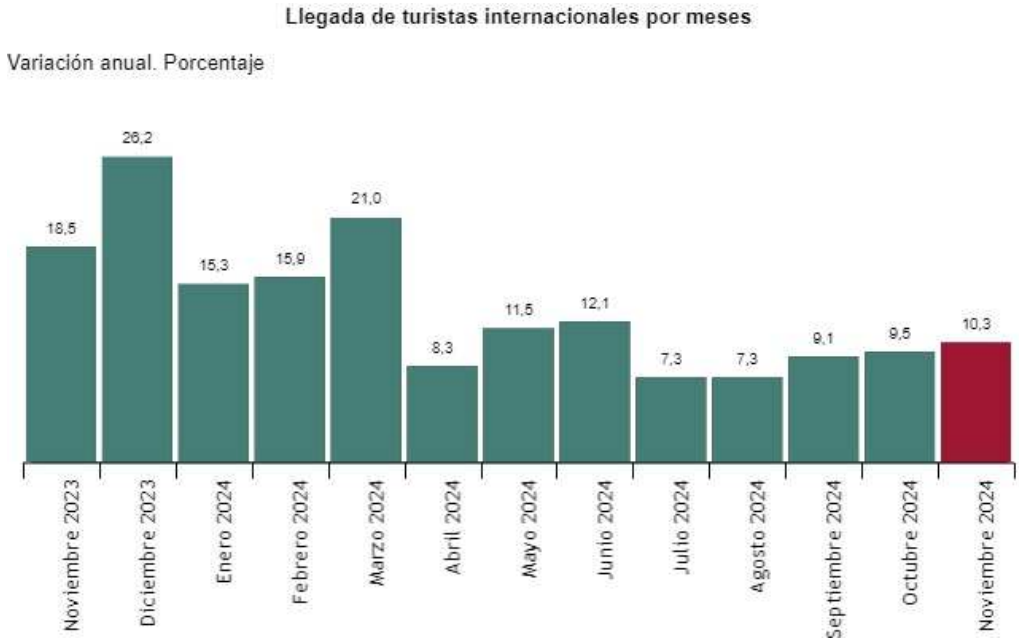

Figura 3. Gráfica de llegada de turistas internacionales por meses. Fuente: (INE, 2024)

En particular, el número de turistas que utilizaron alojamiento hotelero aumentó un 5,7% como se puede apreciar en la Figura 4 [5], y el número de ellos que se decantaron por la vivienda en alquiler (como Airbnb) un 24,5%. ¿Son estos datos motivo de que el precio del suelo urbano crezca?

Llegada de turistas internacionales según tipo de alojamiento. Noviembre 2024

		Datos mensuales		Datos acumulados	
		Valor absoluto	Variación anual (%)	Valor absoluto	Variación anual (%)
TOTAL		5.665.204	10,3	88.533.560	10,7
Alojamiento de mercado	Total alojamiento de mercado	4.356.423	8,0	73.379.364	10,5
	Alojamiento hotelero	3.324.639	5,7	58.705.218	8,2
	Vivienda de alquiler	801.563	24,5	10.904.426	27,3
	Resto alojamiento de mercado	230.222	-5,9	3.769.721	4,9
Alojamiento de no mercado	Alojamiento de no mercado	1.308.781	18,8	15.154.196	11,8
	Vivienda en propiedad	385.026	16,6	4.627.288	5,1
	Vivienda de familiares o amigos	809.850	15,0	9.417.646	13,3
	Resto Alojamiento no de mercado	113.905	68,5	1.109.261	31,2

Figura 4. Llegada de turistas internacionales según tipo de alojamiento. Fuente: (INE, 2024)

Claramente, este flujo turístico genera un impacto significativo en diversos sectores económicos y sociales, incluyendo el mercado inmobiliario y el desarrollo urbano. Por ello, este estudio busca comprender las dinámicas entre el turismo y el mercado del suelo urbano, identificando correlaciones significativas y aplicando modelos que permitan apreciar estas relaciones. También se buscará responder a las preguntas de cuáles son las provincias más afectadas por dichas correlaciones y si estas dinámicas deberían tenerse en cuenta para las políticas de vivienda y uso de suelo, ya que los resultados podrían ser de gran utilidad para la toma de decisiones estratégicas en los ámbitos del turismo, el desarrollo urbano y el mercado inmobiliario, fomentando así un equilibrio entre la sostenibilidad urbana y el desarrollo turístico.

1.1 Objetivos

El objetivo general es analizar la relación entre el flujo turístico y las variaciones en el precio del suelo urbano en España durante el periodo comprendido entre enero de 2017 y junio de 2024.

Entre los objetivos específicos de este proyecto se encuentran:

- Identificar patrones de correlación entre las llegadas de turistas y las fluctuaciones en el precio medio del metro cuadrado del suelo urbano.

- Determinar las provincias donde estas relaciones sean más significativas y analizar los factores contextuales que influyen en dichas tendencias.
- Aplicar métodos estadísticos y de aprendizaje automático (clustering en este caso) para modelar las relaciones entre las variables estudiadas.
- Así mismo, como objetivo añadido se creará un *dashboard* con gráficas ilustrativas que ayuden a la toma de decisiones en ámbitos como, por ejemplo, el entorno inmobiliario, el desarrollo urbano…

1.2 Resultados esperados

Se espera obtener un análisis detallado que documente las correlaciones y relaciones identificadas entre los flujos turísticos y los precios del suelo urbano, aportando gráficos y mapas generados mediante herramientas como R o Python, que permitirán una interpretación visual intuitiva de los resultados.

De igual manera como ya se dijo se creará un *dashboard* interactivo en PowerBI que facilite la exploración de datos y resultados por parte de los interesados. Permitiendo así una toma de decisiones estratégicas informada sobre políticas urbanas, planeación territorial y desarrollo inmobiliario.

1.3 Requisitos del proyecto

Para garantizar el éxito del proyecto, se han definido los siguientes requisitos:

- **Acceso a Bases de Datos Fiables y Oficiales:**

  La base de datos elegida para representar los datos de viajeros y pernoctaciones se obtendrá del Instituto Nacional de Estadística (INE), que ofrece registros mensuales detallados de la ocupación hotelera por provincias [2].

  Por otro lado, los datos sobre el precio medio del metro cuadrado de suelo urbano provendrán del Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA), que publica informes trimestrales [1]. La elección de esta base de datos limita significativamente este estudio en el sentido de que la precisión de las comparaciones entre ambas variables tendrá que ser trimestral en lugar de mensual como permiten los datos de INE.

- **Acotación Temporal:**

  Los datos abarcarán el período de enero de 2017 a junio de 2024 para garantizar consistencia y comparabilidad. Pese a que se tienen datos del precio medio del metro cuadrado de suelo urbano desde 2004 (hasta el segundo trimestre de 2024), como la base de datos de la ocupación hotelera abarca desde enero de 2017 (hasta octubre de 2024), la acotación temporal de este estudio se ve reducida y limitada a la intersección de ambos periodos.

- **Procesos de Limpieza y Normalización:**

  Otro requisito del proyecto es la realización de procesos de limpieza, normalización y análisis exploratorio de los datos para garantizar su consistencia y calidad. Por tanto, los conjuntos de datos serán procesados para corregir inconsistencias, manejar valores atípicos y unificar las unidades de medida.

- **Recursos Humanos y Tecnológicos:**

  El equipo humano necesario y componente de este proyecto es un grupo multidisciplinar con experiencia en análisis de datos, estadística, aprendizaje automático y visualización, formado por un ingeniero informático, una física, un ingeniero de organización industrial y dos matemáticos.

  En cuanto a Hardware, es imprescindible el uso de dispositivos con capacidad suficiente para procesar grandes volúmenes de datos.

  Dentro del Software se requiere el uso de herramientas de análisis y modelado como Python o R (con librerías como caret, cluster, ggplot…), sistemas de gestión de bases de datos SQL, OpenRefine para la curación de datos, y PowerBI para la creación del *dashboard*. Así mismo, para la hacer el diagrama Gantt de planificación temporal se utilizará el software Project Libre.

- **Estructura de Almacenamiento:**

  Se precisa de un sistema seguro y eficiente para almacenar las bases de datos en formatos compatibles como CSV o SQL, como por ejemplo el servicio Cloud de la Universidad de Cantabria (OneDrive).

1.4 Requerimientos Técnicos y Consideraciones.

Para desarrollar el proyecto de manera adecuada y obtener los resultados esperados, se deberán considerar los siguientes aspectos:

- **Cobertura Temporal y Geográfica.** Durante un preprocesado y curación de los datos, será necesario asegurar que ambos conjuntos de datos cubran el periodo de estudio (enero de 2017 - junio de 2024) y establecer una cobertura geográfica a nivel provincial que permita identificar diferencias contextuales entre regiones, de manera completa y precisa.
- **Homogeneización de Datos:** normalizar las unidades de medida y formatos de ambos conjuntos de datos para facilitar su integración y comparación.
- **Seguridad y Confidencialidad:** garantizar el manejo seguro y ético de los datos durante todas las etapas del proyecto.
- **Documentación Exhaustiva:** elaborar documentación detallada, incluyendo un Plan de Gestión de Datos (DMP), diagramas de Gantt para la planificación, y repositorios de metadatos que faciliten la reproducción de resultados.

Con este enfoque estructurado y los recursos necesarios, y asegurándose en todo momento la disponibilidad de los datos en el presente y futuro, se espera que el proyecto proporcione información valiosa para optimizar la gestión turística y urbana en España.

2 Descripción de las fuentes de datos:

2.1 Orígenes

Las bases de datos empleadas en este proyecto provienen de fuentes oficiales y confiables como son el Instituto Nacional de Estadística (INE) y el Ministerio de Transportes, Movilidad y Agenda Urbana (MITMA) [1] [2].

El INE ha proporcionado una base de datos mensual sobre la capacidad hotelera, incluyen un desglose por provincia. Por su lado, el documento del Ministerio ofrece datos trimestrales sobre el precio medio del metro cuadrado de suelo urbano, también con desglose provincial.

2.2 Formatos

Ambas bases de datos se encuentran disponibles en formato Excel accesibles desde los respectivos portales web oficiales de ambas instituciones. Ambos formatos son compatibles con las herramientas de análisis seleccionadas (Python, R, OpenRefine…), facilitando la integración, limpieza y análisis de los datos. Además, se implementarán procesos de transformación para homogeneizar los formatos y asegurar la interoperabilidad entre los conjuntos de datos.

2.3 Cobertura

Ambas bases de datos dan datos a nivel tanto provincial como a nivel autonómico de España. Por lo que, en cuanto a la cobertura geográfica, el análisis se llevará a cabo a nivel provincial en España. Este enfoque permitirá observar patrones regionales y diferencias contextuales entre provincias con alta afluencia turística, como las Islas Baleares y Barcelona, frente a otras con menor actividad turística.

Por otro lado, dado que ambas bases de datos no coinciden exactamente en su acotación temporal como se ha mencionado anteriormente, la cobertura temporal se ha visto altamente condicionada por dicha disponibilidad de datos. De este modo, el intervalo de tiempo elegido para la realización de este estudio es el periodo abarcado entre enero de 2017 y junio de 2024, de manera trimestral. Este intervalo se cree suficiente para poder observar tendencias a medio y largo plazo, incluyendo posibles impactos de eventos globales como la pandemia de COVID-19 y la recuperación posterior.

A pesar de estas limitaciones, el enfoque provincial garantiza un análisis robusto que puede generar información valiosa sobre las tendencias nacionales y regionales.

3 Diagrama Gantt

Una vez realizada la descripción general del proyecto y de los datos a utilizar, es conveniente mostrar cuál sería la distribución temporal del proyecto. En este caso, esta distribución engloba todo el ciclo de vida de los datos, desde la planificación y adquisición de estos hasta el análisis y creación de gráficos para estudiar el impacto del turismo en España en el precio del suelo urbano, que es el principal acometido de este proyecto.

En este caso para realizar una correcta planificación del proyecto se dividieron las tareas en distintos paquetes de trabajo, donde dentro de cada uno de ellos se definieron distintos niveles con subtareas a las que se les asignó una duración determinada estimada, estableciendo a su vez las distintas relaciones de precedencia con las otras subtareas. En este caso los paquetes de trabajo en los que se dividieron las tareas fueron:

1. Definición y planificación del proyecto
2. Recolección y Generación Bases de Datos
3. Curación y Procesamiento de las Bases de Datos
4. Análisis Exploratorio de Datos (EDA)
5. Modelado y Análisis Avanzado
6. Diseño y Creación del Dashboard
7. Revisión final proyecto y presentación

Estos paquetes de trabajo podrán ser analizados más en detalle en el informe Gantt adjunto al trabajo, generado en este caso utilizando el software Project Libre. En este informe se muestran estos paquetes de tareas, junto a los distintos niveles de subtareas en una tabla, donde además figura en otras páginas el diagrama Gantt generado a partir de estas tareas. Dentro de las tareas se podrán apreciar tareas con duración cero que en este caso hacen referencia a los distintos hitos del horizonte temporal, necesarios para ver si se van cumpliendo con éxito las distintas metas del proyecto.

Ha de notarse que dada la gran dimensión del diagrama Gantt no es viable mostrarlo por completo en una sola página. Es por ello que el software de Project Libre lo que hace es mostrar primero la tabla con las tareas y después mostrar en las siguientes páginas el diagrama Gantt asociado a estas tareas.

En este caso el documento relativo al diagrama de Gantt se muestra por separado en otro documento dada su gran extensión.

4 Presupuesto

Para poder realizar una estimación realista de los costes, es necesario tener una estimación de las horas totales de trabajo por cada paquete de trabajo (mostradas en el informe del Gantt) y multiplicarlas por una tarifa promedio por hora que incluya además de la

experiencia del equipo involucrado, los costes asociados a los activos fijos (infraestructura de almacenamiento, procesamiento, análisis…). Para este proyecto, se asumirá una tarifa promedio de 50 € por hora para un proyecto de nivel intermedio, que incluye los costes comentados anteriormente. Coma cada día consta de 8 horas de trabajo, los cálculos se harían de la siguiente forma:

1. **Definición y Planificación del Proyecto (55 días)**
   Horas totales: 55 días * 8 horas/día = 440 horas.
   Coste estimado: 440 horas * 50 €/hora = 22,000 €.
2. **Recolección y Generación de Bases de Datos**
   - Recolección Datos Ocupación Hotelera:
     Total de días: 232 días * 8 horas/día = 1,856 horas
     Coste estimado: 1,856 horas * 50 €/hora = 92,800 €
   - Recolección Datos Precio Suelo Urbano:
     Total de días: 216 días * 8 horas/día = 1,728 horas
     Coste estimado: 1,728 horas * 50 €/hora = 86,400 €

   Subtotal: 179,200 €

3. **Curación y Procesamiento de las Bases de Datos (55 días)**
   Horas totales: 55 días * 8 horas/día = 440 horas.
   Coste estimado: 440 horas * 50 €/hora = 22,000 €.
4. **Análisis exploratorio de Datos (EDA) (33 días)**
   Horas totales: 33 días * 8 horas/día = 264 horas.
   Coste estimado: 264 horas * 50 €/hora = 13,200 €.
5. **Modelado y Análisis Avanzado (9 días)**
   Horas totales: 9 días * 8 horas/día = 72 horas.
   Coste estimado: 72 horas * 50 €/hora = 3,600 €.
6. **Diseño y Creación del Dashboard (10 días)**
   Horas totales: 10 días * 8 horas/día = 80 horas.
   Coste estimado: 80 horas * 50 €/hora = 4,000 €.
7. **Revisión Final y Presentación (7 días)**
   Horas totales: 7 días * 8 horas/día = 56 horas.
   Coste estimado: 56 horas * 50 €/hora = 2,800 €.

COSTE TOTAL ESTIMADO: 246,800 €

Como se aprecia, en base a estos cálculos se obtiene un precio estimado de este proyecto (asumiendo creación desde cero) de 246,800 €. Nótese que para obtener este precio se tuvo que hacer una estimación muy simplificada y asumiendo ciertas hipótesis. Para obtener un coste más detallado habría que hacer una imputación de las horas hombre y horas de los activos fijos para cada subtarea dentro de cada paquete de trabajo y establecer los salarios para cada caso. No obstante, dada la dificultad para obtener este nivel de detalle se decidieron hacer estas asunciones.

5 Curación de las bases de datos

En este apartado se van a comentar las técnicas empleadas para curar las bases de datos utilizadas en este proyecto, así como aquellas que pudieron utilizar tanto el INE como el Ministerio para ponerlas posteriormente a disposición de la ciudadanía.

Comenzando en primer lugar con las técnicas utilizadas por cuenta propia sobre las bases de datos originales, se encuentran las siguientes:

1. **Ajuste del margen temporal**

   Como ya se comentó anteriormente, ambas bases de datos contenían información con diferentes rangos temporales y periodicidades:

   - Ocupación hotelera: datos mensuales desde enero de 2017 a octubre de 2024.
   - Precio del suelo urbano: datos trimestrales desde 2004 hasta el segundo trimestre de 2024.

   En este caso para hacer los datos comparables, se ajustaron ambos conjuntos de datos al rango temporal común (enero 2017 a junio de 2024) y transformando la periodicidad de los datos mensuales en trimestrales para garantizar la coherencia de ambos *datasets*.

2. **Simplificación de datos geográficos**

   Ambos conjuntos de datos contenían además de filas relativas a cada provincia, filas relativas también a cada comunidad autónoma. En este caso como el análisis efectuado en este proyecto no requería desglose a nivel de comunidad autónoma, se decidieron eliminar estas filas para simplificar el conjunto de datos. No obstante,

aunque esta decisión no era estrictamente necesaria, reduce la complejidad y mejora la calidad del análisis.

3. **Gestión de valores faltantes**

   Para abordar la presencia de valores faltantes en los datos se decidió utilizar una técnica de interpolación lineal. Este método permite estimar los valores ausentes calculando la media de los datos inmediatamente anterior y posterior dentro de la misma provincia.

4. **Eliminación de *outliers***

   Para garantizar que los datos fueran representativos y evitar distorsiones en el análisis, se llevó un proceso de identificación y eliminación de valores atípicos. Este paso se realizó mediante la detección de puntos de datos que se desvían significativamente de las tendencias generales, utilizando técnicas estadísticas como el rango intercuartílico o el análisis visual mediante gráficos de dispersión.

5. **Formato y exportación**

   En este caso, los datos estaban al principio en formato Excel (.xlsx) por lo que se exportaron a archivos CSV (.csv) para facilitar su manejo con herramientas de análisis como R. Durante este proceso, se identificaron varios problemas que se resolvieron con el software de OpenRefine visto en la asignatura:

   - Comillas innecesarias: en este caso los datos estaban rodeados de comillas, lo que impedía que R los detectara como valores numéricos. Los nombres de las columnas también estaban entre comillas, lo que hacía que R identificara únicamente una columna en el archivo. Para solucionarlo, se decidió eliminar las comillas y dividir correctamente las columnas.
   - Formato de valores: en lo que respecta a los valores numéricos, estos no se reconocían como tal. Aunque este problema podía solucionarse directamente con R, se decidió utilizar de nuevo OpenRefine para evitar posibles errores posteriores.

Todos estos cambios sobre las bases de datos fueron realizados durante este proyecto con el fin de poder realizar un análisis correcto y sin ningún tipo de error. No obstante, si bien estos cambios ayudaron a dejar estos *datasets* en perfecto estado para su uso, existen otras posibles técnicas a mencionar que pudieron ser utilizadas por parte del INE y el Ministerio

antes de publicar estas bases de datos. Por mencionar algunas, se pudieron haber utilizado las siguientes:

- **Validación de datos recopilados:** donde se pudieron comparar los datos reportados con fuentes oficiales o independientes para garantizar la consistencia y la fiabilidad. También se podría haber verificado que los datos cumplen con las reglas definidas, como rangos aceptables de valores o formatos específicos.
- **Gestión de valores faltantes:** así como la interpolación lineal, se podrían haber utilizado otras técnicas más avanzadas como imputación basada en algoritmos (e.g. k-nearest neighbors o modelos predictivos como árboles de regresión entre otros).
- **Detección de *outliers*:** seguramente se han utilizado métodos estadísticos robustos, como el análisis de series temporales o modelos econométricos para identificar datos que se desvíen significativamente de las tendencias históricas.
- **Ajustes estacionales:** para los datos relacionados con la ocupación hotelera, es posible que se hayan ajustado los datos para tener en cuenta la estacionalidad, proporcionando series ajustadas que reflejan tendencias subyacentes sin variaciones estacionales.
- **Análisis y corrección de sesgos:** los datos iniciales recopilados por estos organismos podrían haber contenido sesgos derivados de errores humanos en el proceso de recolección, como valores duplicados o mal registrados. Una técnica que podría haberse utilizado para este aspecto es el uso de algoritmos de detección de duplicados o inconsistencias, como la comparación de registros por patrones comunes o reglas definidas.

Estas son algunas de las posibles técnicas de curación que han podido haber utilizado estos dos organismos para poner a disposición estas bases de datos. Si bien estas técnicas son útiles y relevantes, seguramente existen muchas más con un mayor nivel de sofisticación que han podido ser utilizadas con el fin de garantizar la correcta estructura e integridad de estos *datasets*.

6 Plan de preservación

En este apartado se detallan los pasos necesarios para garantizar la preservación, accesibilidad y reutilización de los datos relacionados con el proyecto “Impacto del

Turismo en el Precio del Suelo Urbano". Los datos incluyen series temporales y datos geográficos a nivel provincial de España, con una temporalidad entre 2017 y 2024. A continuación, se listarán de manera esquemática los detalles relativos a este plan de preservación:

6.1 Identificación de los datos a preservar

- **Datos originales:**
  - **Fuente:** INE (ocupación hotelera) y MITMA (precio del suelo urbano).
  - **Formato:** XLSX.
  - **Descripción:** Datos a nivel provincial sobre ocupación hotelera mensual y precios medios trimestrales del metro cuadrado de suelo urbano.
- **Datos preprocesados:**
  - **Formato:** CSV.
  - **Descripción:** Datos ajustados, normalizados y unificados para análisis posterior.
- **Resultados analíticos:**
  - **Formato:** PNG (visualizaciones), CSV (resultados de modelos), Power BI (*dashboard*).
  - **Descripción:** Gráficas, modelos de clustering, regresión y resultados de análisis estadísticos.
- **Documentación:**
  - **Formato:** PDF (informe final).
  - **Descripción:** Documentos explicativos del proceso, manuales de usuario y scripts.

6.2 Formatos de archivo

- **Datos tabulares:** CSV (para interoperabilidad y compresión).
- **Visualizaciones:** PNG y Power BI para *dashboards* interactivos.

- **Scripts:** Python (.py) y R (.R).
- **Documentación:** PDF.

6.3 Almacenamiento

- **Temporal (durante el proyecto):**
  - **Ubicación:** OneDrive de la Universidad de Cantabria.
  - **Características:** Acceso controlado para el equipo de trabajo mediante autenticación segura.
- **Largo plazo (post-proyecto):**
  - **Repositorio abierto:** Zenodo o Figshare, que asignan un DOI a los *datasets*.
  - **Licencia:** Creative Commons Attribution (CC-BY) para fomentar la reutilización.

6.4 Procesos de curación

1. **Homogeneización temporal:**
   - Ajustar los datos de ocupación hotelera (mensuales) a intervalos trimestrales para coincidir con los datos del precio del suelo.
2. **Eliminación de duplicados:**
   - Uso de pandas en Python para identificar y eliminar registros duplicados, así como uso también de Open Refine para esta tarea.
3. **Interpolación de valores faltantes:**
   - Aplicar interpolación lineal para estimar valores ausentes.
4. **Detección de *outliers*:**
   - Aplicar el rango intercuartílico para identificar y eliminar valores anómalos usando Open Refine.

5. **Conversión de formatos:**
    - Transformar datos de XLSX a CSV para facilitar el análisis.

6.5 Metadatos

Los metadatos utilizados para el proyecto se han extraído mediante una búsqueda haciendo uso de expresiones regulares en los HTML de las páginas de las bases de datos. De esta manera se ha creado un script Python que busca las palabras clave en el código para poder obtenerlos. Una vez obtenidas las palabras clave, se guardan en un archivo XML en el que aparecen representados los metadatos en un formato DublinCore.

De esta manera, se han obtenido los siguientes metadatos para cada una de las páginas de las bases de datos mostrados en las Figuras 5 y 6:

metadata_viajeros.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="https://www.ine.es/jaxiT3/Tabla.htm?t=49371">
        <dc:title>Viajeros y pernoctaciones por comunidades, ciudades autónomas y provincias(49371)</dc:title>
        <dc:creator>INE - Instituto Nacional de Estadística</dc:creator>
        <dc:publisher>INE - Instituto Nacional de Estadística</dc:publisher>
        <dc:description>INE. Instituto Nacional de Estadística. National Statistics Institute. Spanish Statistical Office. El INE elabora y
        <dc:subject>Comunidades Autónomas y Provincias, Viajeros y pernoctaciones, Residencia, Encuesta de Ocupación Hotelera, EOH</dc:subj
        <dc:date>2017-01-01 - 2024-12-23</dc:date>
        <dc:language>es</dc:language>
        <dc:source>https://www.ine.es</dc:source>
        <dc:type>WebSite</dc:type>
        <dc:format>HTML</dc:format>
        <dc:identifier>49371</dc:identifier>
        <dc:coverage>Total Nacional, 01 Andalucía, 04 Almería, 11 Cádiz, 14 Córdoba, 18 Granada, 21 Huelva, 23 Jaén, 29 Málaga, 41 Sevilla,
        <dc:rights>https://www.ine.es/dyngs/AYU/index.htm?cid=125</dc:rights>
    </rdf:Description>
</rdf:RDF>
```

Figura 5. Metadatos de la base de datos de ocupación hotelera. Fuente: Elaboración propia

metadata_boletin.xml

```xml
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=36000000">
        <dc:title>Boletín estadístico online - Información estadística - Ministerio de Fomento</dc:title>
        <dc:creator>Transportes y Movilidad Sostenible</dc:creator>
        <dc:publisher>Transportes y Movilidad Sostenible</dc:publisher>
        <dc:description>Sin descripción</dc:description>
        <dc:subject>Categorías: Transacciones de suelo, Valor de las transacciones de suelo, Superficie de las transacciones de suelo, Prec
        <dc:subject>Subcategorías: 1. Número total de transacciones de suelo por comunidades autónomas y provincias, 1.1. Transacciones de
        <dc:language>es</dc:language>
        <dc:source>https://apps.fomento.gob.es/BoletinOnline2/</dc:source>
        <dc:type>WebSite</dc:type>
        <dc:format>HTML</dc:format>
        <dc:identifier>https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=36000000 --></dc:identifier>
        <dc:coverage>Sin cobertura específica</dc:coverage>
        <dc:rights>https://www.transportes.gob.es/ministerio/aviso-legal</dc:rights>
    </rdf:Description>
</rdf:RDF>
```

Figura 6. Metadatos de la base de datos de precio del suelo urbano. Fuente: Elaboración propia

6.6 Acceso y reutilización

- **Acceso durante el proyecto:** Limitado al equipo, mediante control de roles.
- **Acceso post-proyecto:** Público mediante repositorios abiertos (Zenodo).
- **Licencia:** CC-BY, permitiendo redistribución con atribución.

6.7 Seguridad

- **Cifrado:** Archivos cifrados con AES-256 durante la transferencia y almacenamiento.
- **Backups:** Copias de seguridad automáticas semanales en Google Cloud.

6.8 Revisión

- **Cronograma:** Revisión de los datos cada dos años para asegurar su accesibilidad y formato actualizado.

6.9 Roles del equipo

- **Gestor del proyecto:** Supervisión general.
- **Curadores de datos:** Procesamiento y ajuste de los datos.
- **Analistas:** Generación de resultados y visualizaciones.
- **Administrador técnico:** Gestión de infraestructura y repositorios.

7 Análisis de los datos

Como ya se comentó al principio, el principal objetivo de este proyecto es estudiar el impacto que ha tenido el flujo turístico en España sobre el precio del metro cuadrado de suelo urbano. De cara a estudiar este impacto utilizando las dos bases de datos ya comentadas, es de gran utilidad el uso de diferentes gráficas que ayuden a ilustrar este análisis.

En primer lugar, resulta pertinente estudiar si existe alguna relación directa entre el número de plazas hoteleras y el precio del suelo. Para llevar a cabo este análisis, se decidió calcular la media por provincia de las variables presentes en ambos conjuntos de datos.

Este enfoque permite sintetizar la información y facilita la identificación de patrones generales.

Como se aprecia en la gráfica de la Figura 7, se pone de manifiesto la relación creciente existente entre ambas variables. Según la teoría, estas variables deberían ser independientes, lo cual se reflejaría en una línea paralela al eje x. No obstante, la inclinación observada en la línea de tendencia indica lo contrario. En base a esta gráfica se puede llegar a la conclusión preliminar de que como es de esperar, un mayor número de turistas está asociado con un incremento en el precio del suelo urbano de las provincias analizadas. Esto se puede deber, por ejemplo, a que estas provincias gozan de una calidad de vida muy buena ya sea dada por los numerosos servicios a los que puede acceder la población o su localización privilegiada que presenta un clima favorable.

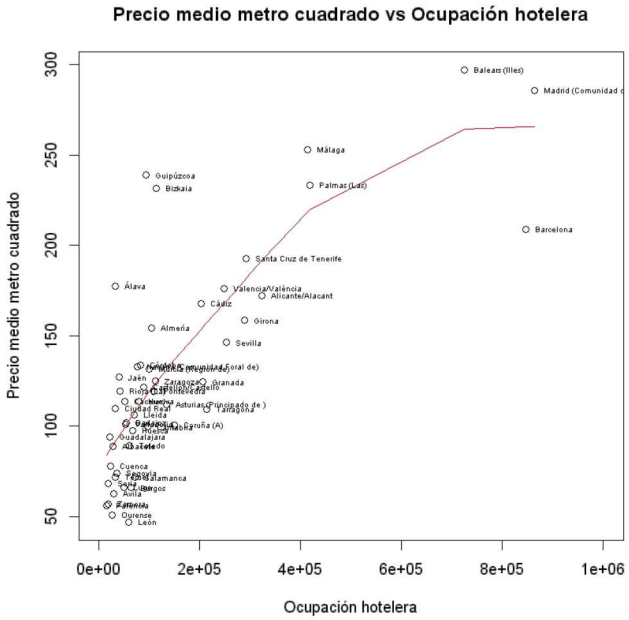

Figura 7. Gráfico de relación entre el precio del suelo urbano y la ocupación hotelera. Fuente: Elaboración propia

Por otro lado, es relevante evaluar si esta relación ha evolucionado a lo largo del tiempo. Para ello, se decidió hacer una comparación entre los datos correspondientes al primer año completo del análisis (con el fin de evitar distorsiones estacionales) y los datos recopilados en el segundo semestre de 2023 junto con el primero de 2024. Esta comparación se muestra en la Figura 8.

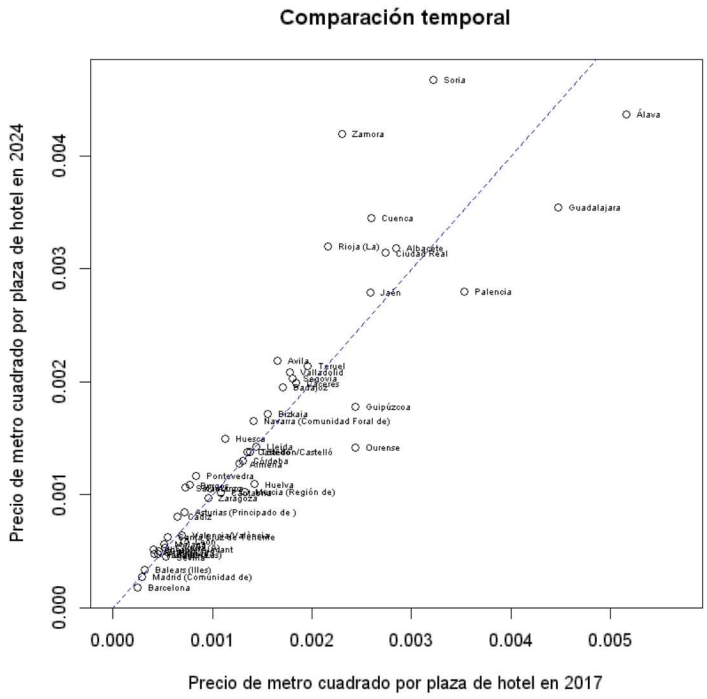

Figura 8. Gráfico de evolución de la relación entre las variables. Fuente: Elaboración propia

Como se aprecia en la Figura 8, los valores situados por encima de la línea discontinua en la gráfica indican que el precio del suelo en el período reciente es superior al registrado en 2017, mientras que aquellos por debajo de la línea reflejan una reducción en el precio del suelo respecto al mismo año. No obstante, los resultados no revelan una tendencia clara en ninguna dirección. Por lo tanto, se puede decir que no se observa ninguna evolución temporal significativa en la relación entre el número de plazas hoteleres y precio del suelo urbano.

Los análisis anteriores han sido de gran utilidad de cara a estudiar la relación entre las variables. Sin embargo, aún se puede ir algo más allá y aplicar algún método de *Machine Learning* no supervisado que permita hacer grupos de provincias (clústering) en base a estas dos variables. Nótese que la aplicación de algún método avanzado de este estilo quedaba reflejada como una de las tareas dentro del diagrama Gantt. En este caso con el fin de obtener estos clústeres de provincias, se decidió utilizar el método de las k-medias, que coge como variables la ocupación hotelera y el precio del suelo urbano. En la Figura 9 se muestra un diagrama de dispersión con los clústeres obtenidos tras aplicar este método.

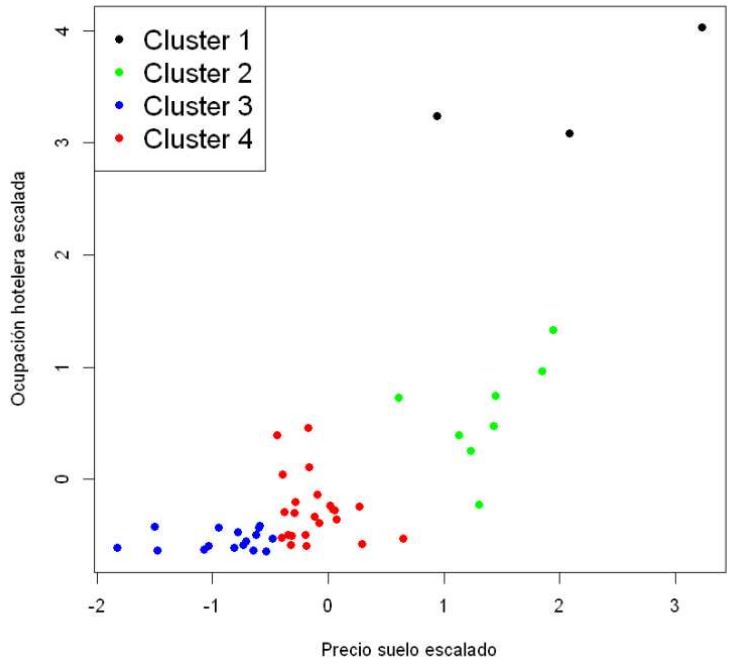

Figura 9. Gráfico de clústeres de provincias. Fuente: Elaboración propia

Como se aprecia en la Figura 9, los resultados de este algoritmo de *Machine Learning* arrojaron un total de 4 clústeres de provincias distinguidos por colores. En primer lugar, se observan los puntos negros del clúster 1 relativos a provincias que presentan una alta ocupación hotelera y un precio alto del suelo urbano. Por otro lado, se distinguen las provincias en color verde del clúster 2 donde si bien el precio del suelo urbano es casi igual de alto al de las del clúster 1, la ocupación hotelera es significativamente inferior. Luego, se encuentra el clúster 3 distinguido en azul donde se encuentran las provincias con menor ocupación hotelera y menor precio del suelo urbano. Por último, se encuentran distinguidas en rojo las provincias relativas al clúster 4 que presentan un precio del suelo urbano mayor a las del clúster 3 y generalmente (salvo en 4 observaciones) una ocupación hotelera similar a las de este clúster. Esta distinción entre estos 4 clústeres permite diferenciar aquellas provincias con alto flujo turístico y precio del suelo urbano, de aquellas que son menos turísticas y que por ende suelen presentar, como ya se vio en la relación lineal de la Figura 7, un menor coste del precio del suelo urbano.

De cara a saber cuáles son las provincias dentro de estos 4 clústeres, se decidió hacer un dendrograma mostrado en la Figura 10.

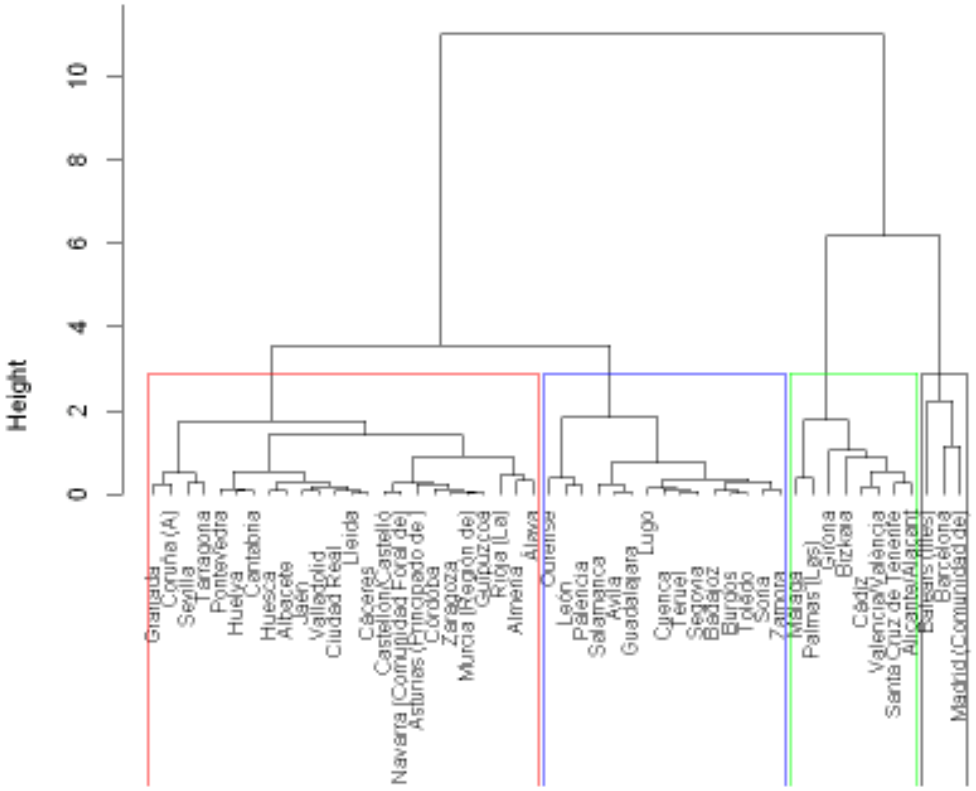

Figura 10. Dendograma de clústeres de provincias. Fuente: Elaboración propia

En este dendrograma se aprecia que las 3 provincias que generalmente presentan mayor flujo turístico y son más caras para vivir, son las Islas Baleares, Barcelona y Madrid. Este resultado era bastante esperable. Por otro lado, se distingue en verde el clúster 2 de las provincias en el segundo puesto en cuanto a flujo turístico y coste del suelo urbano donde se sitúan por ejemplo Málaga, Las Palmas de Gran Canaria, Girona… Después, se observa el clúster 4 en color rojo que es el que más provincias comprende donde están aquellas que no presentan un flujo turístico tan elevado como las de los clústeres 1 y 2, y que tampoco son excesivamente caras para vivir. Algunas de ellas son, por ejemplo, Asturias, Cantabria, Almería, A Coruña, Valladolid… Por último, representadas en azul están las provincias del clúster 3 relativas a aquellas donde hay poco flujo turístico y el precio del suelo urbano es menor donde se encuentran por ejemplo Palencia, Ourense, Salamanca, Ávila…

Como conclusión se puede decir que los resultados de este proyecto han revelado que sí que parece existir cierta correlación durante los últimos años entre el flujo turístico y el precio del suelo urbano en las distintas provincias de España, donde como se vio aquellas

con mayor turismo presentan un coste relativo al precio del suelo urbano significativamente alto. Este aspecto puede deberse a que un mayor turismo en una provincia se puede traducir en que exista un ratio de hoteles (o apartamentos turísticos) por viviendas propias superior al de otras provincias con menos turismo, lo cual hace que al haber menos disponibilidad de vivienda propia se encarezcan los precios de suelo urbano. A esto también se le suma el elevado precio base del que se parte por la localización de la provincia.

Con esto se puede ver que los objetivos del proyecto relativos al estudio de estas variables se consiguieron alcanzar satisfactoriamente. No obstante, todavía queda por abordar un último objetivo del proyecto citado en los primeros apartados, relativo al desarrollo de un cuadro de mandos (o *dashboard*) que pueda ser de utilidad para la toma de decisiones de empresas ligadas a ámbitos del turismo, desarrollo urbano, mercado inmobiliario… En este caso, utilizando el software de Power Bi se creó un sencillo cuadro de mandos donde quien quiera puede visualizar ciertas métricas relativas a estas variables filtrando por provincias y accediendo a su vez a gráficas interactivas con datos sobre la ocupación hotelera y precio del suelo urbano. En la Figura 11 se muestra una captura de este cuadro de mandos que se mostró en la presentación del proyecto del día 17 de enero. Así mismo, se pone a disposición un link para poder interactuar con este cuadro de mandos: https://app.powerbi.com/view?r=eyJrIjoiYzUyMjY3M2ItNjQ3Yi00OGZlLWEwNjgtYjQ5Mzg0ZmE2MGRlIiwidCI6IjA1ZWE3NGEzLTkyYzUtNGMzMS05NzhhLTkyNWMzYzc5OWNkMCIsImMiOjh9

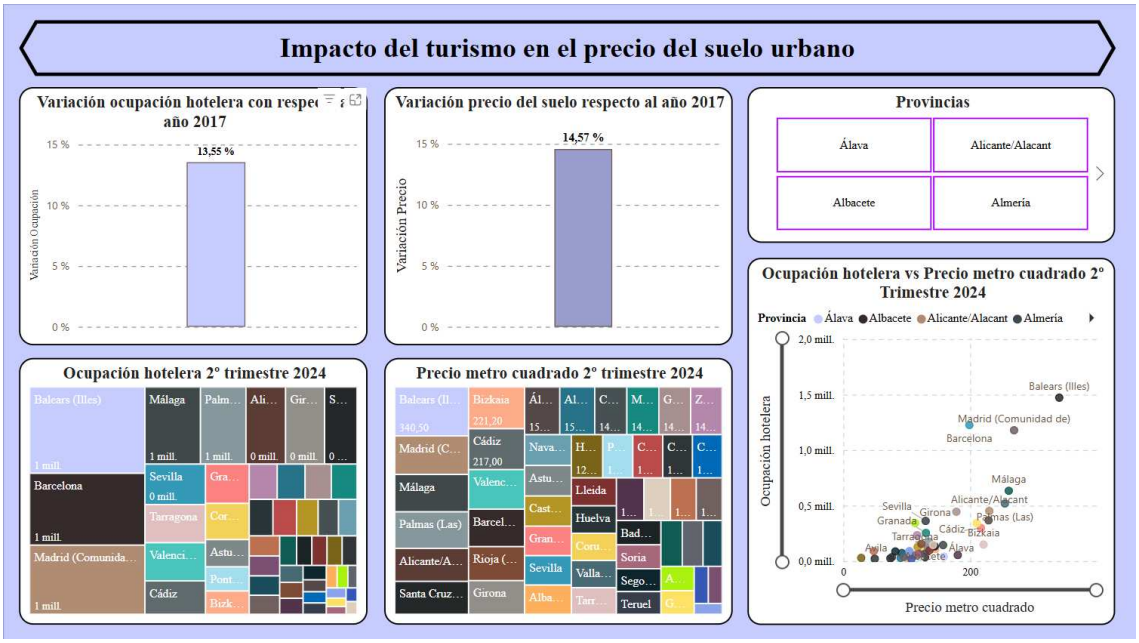

Figura 11. Captura del *dashboard* del impacto del turismo en el precio del suelo urbano. Fuente: Elaboración propia

8 Bibliografía

[1] «Boletín estadístico online - Información estadística - Ministerio de Fomento», Ministerio de Transportes y Movilidad Sostenible. Accedido: 10 de enero de 2025. [En línea]. Disponible en: https://apps.fomento.gob.es/BoletinOnline2/?nivel=2&orden=36000000

[2] «Viajeros y pernoctaciones por comunidades, ciudades autónomas y provincias(49371)», INE. Accedido: 10 de enero de 2025. [En línea]. Disponible en: https://www.ine.es/jaxiT3/Tabla.htm?t=49371

[3] «Estadísticas de turismo en España en 2023 - ¿Cuántas visitas recibe?», Road Genius. Accedido: 10 de enero de 2025. [En línea]. Disponible en: https://roadgenius.com/es/statistics/tourism/spain/

[4] «Nota de Prensa: Estadística de Movimientos Turísticos en Fronteras (FRONTUR). Noviembre 2024. Datos provisionales.», INE. Accedido: 10 de enero de 2025. [En línea]. Disponible en: https://www.ine.es/dyngs/Prensa/es/FRONTUR1124.htm

[5] «Nota de Prensa: Encuesta de Turismo de Residentes (ETR/FAMILITUR). Cuarto trimestre 2023 y año 2023.», INE. Accedido: 10 de enero de 2025. [En línea]. Disponible en: https://ine.es/dyngs/Prensa/ETR4T23.htm