Facultade de Informática
UNIVERSIDADE DA CORUÑA

TRABAJO FIN DE GRADO
GRAO EN CIENCIA E ENXEÑARÍA DE DATOS

Herramienta de análisis y explotación de datos turísticos

Estudiante: Andrea García Álvarez

Dirección: Ana B. Cerdeira Pena

Guillermo de Bernardo Roca

A Coruña, Septiembre de 2023

A mi madre, siempre vas a estar conmigo

Agradecimientos

Gracias a mi padre y a mi hermana por apoyarme y soportarme estos últimos 4 años y a lo largo de mi vida.

También quiero agradecer a mis amigos de siempre y a los nuevos incorporados en esta etapa tan bonita, aquí empieza una nueva.

A cada una de las personas que me ha acompañado, acompaña y acompañará en mi camino.

Por último, no me olvido de la dedicación de mis dos tutores, Ana y Guillermo, a lo largo de este proyecto, gracias por su infinita paciencia y orientación.

Resumen

El objetivo de este trabajo de fin de grado es el desarrollo de procesos para la extracción y exploración de datos procedentes de la plataforma Airbnb y del Instituto Nacional de Estadística vinculados al sector del turismo en España, con especial atención al mercado de la oferta vacacional, y la creación de *dashboards* interactivos para el posterior análisis y explotación de los mismos, que permitan obtener conocimiento derivado de utilidad.

Para alcanzar dicho objetivo hemos dividido el desarrollo en fases. En la primera de ellas definimos el alcance del proyecto y sus objetivos principales, seguido de un estudio de aplicaciones y tecnologías que facilitasen su desarrollo. A continuación, se realizó un primer estudio de los datos a tratar y se comenzó con el desarrollo iterativo del trabajo, incluyendo, entre otros, procesos de extracción y limpieza, diseño de un almacén de datos, análisis, visualización y extracción de conclusiones.

Para la realización de este trabajo se ha utilizado el lenguaje Python, junto con las librerías *BeautifulSoup* y *Selenium* para la extracción de datos de la Web. Además, hemos hecho uso de MySQL y DBeaver para la creación de un almacén de datos que cumpliese con nuestras necesidades. Por último, para la creación de los diferentes *dashboards* y gráficas acudimos a la herramienta de Power BI.

El proyecto se ha realizado siguiendo una metodología basada en iteraciones, estando en constante contacto con los directores y llevando a cabo reuniones de seguimiento.

Abstract

The objective of this end-of-degree project is the development of processes for the extraction and exploration of data from the Airbnb platform and the National Institute of Statistics linked to the tourism sector in Spain, with special attention to the market for vacation offers, and the creation of interactive dashboards for subsequent analysis and exploitation of the same, which allow obtaining knowledge derived from utility.

To achieve this objective we have divided the development into phases. In the first of them we defined the scope of the project and its main objectives, followed by a study of applications and technologies that would facilitate its development. Next, a first study of the data to be processed was carried out and the iterative development of the work began, including, among others, extraction and cleaning processes, design of a data warehouse, analysis, visualization and drawing of conclusions.

During the execution of this work, the Python language was used along with the *BeautifulSoup* and *Selenium* libraries to extract data from the Web. In addition, we have made use

of MySQL and DBeaver to create a data warehouse that meets our needs. Finally, to create the different dashboards and graphs we used the Power BI tool.

The project has been carried out following a methodology based on iterations, being in constant contact with the directors and holding follow-up meetings.

Palabras clave:

- Airbnb
- INE
- Python
- Almacén de datos
- ETL
- MySQL
- PowerBi
- Dashboards

Keywords:

- Airbnb
- INE
- Python
- Data Warehouse
- ETL
- MySQL
- PowerBi
- Dashboards

Índice general

Índice de figuras

Índice de tablas

Capítulo 1
Introducción

En este capítulo se presenta la motivación del proyecto, seguida por sus principales objetivos y la estructura de la presente memoria.

1.1 Motivación

En la actualidad España es un destino turístico líder a nivel mundial, lo que ha convertido al turismo en un sector clave en el crecimiento económico del país, generando importantes ingresos y empleo.

Disponer de herramientas que permitan realizar estudios de análisis de características y descubrimiento de tendencias en el ámbito turístico, así como ofrecer una visión general de la oferta/demanda del sector devienen de gran interés dada la relevancia manifiesta del contexto. Esto ha dado pie a la aparición de muchos estudios y herramientas web de visualización de datos, como pueden ser por ejemplo las del Instituto Nacional de Estadística (INE), que nos permiten conocer datos generales como los gastos por viaje, duración, principales destinos, etc.

Sin embargo, la gran mayoría de estos estudios se limitan a la información recogida en diferentes encuestas a la población y no ofrecen ninguna visión proporcionada directamente de las diferentes aplicaciones y plataformas de oferta turística utilizadas hoy en día para la reserva de alojamientos.

En nuestro país Airbnb es una de las plataformas líderes en ofertas de alojamiento. En los últimos años ha experimentado un crecimiento significativo en España, convirtiéndose en una fuente valiosa de información. En este contexto, el objetivo de este proyecto es integrar la información recopilada en encuestas con los datos obtenidos a través de la plataforma Airbnb para obtener una visión más completa y detallada del mercado turístico en el país.

Estos datos permitirán a investigadores, planificadores turísticos y empresas del sector tomar decisiones más informadas y estratégicas.

1.2 Objetivos

La realización de este proyecto se puede dividir en tres grandes bloques de trabajo. El primero de ellos consiste en extraer los datos de Airbnb y otras fuentes relevantes, y realizar una limpieza inicial para asegurar la calidad y consistencia de los datos obtenidos. El segundo bloque se centra en la creación de una base de datos que integre los datos recopilados completando así el proceso Extraction, Transformation and Load (ETL). El tercer bloque se basa en el análisis de estos datos y la creación de visualizaciones informativas significativas.

En este contexto, se plantean los siguientes objetivos:

- Desarrollo de procesos para la extracción y limpieza de datos vinculados al sector turístico procedentes de diversas fuentes. Se utilizarán técnicas de *web scraping* para obtener información relevante directamente de la plataforma de Airbnb. Esto implicará la selección de variables de interés y la aplicación de técnicas de limpieza y preprocesamiento para asegurar la calidad de los datos. Además, se considerará la integración de datos procedentes del Instituto Nacional de Estadística para enriquecer la información recopilada.

- Diseño de un almacén de datos para la integración de los datos obtenidos. Se creará una base de datos que permita la integración de los datos obtenidos anteriormente. Se definirán las tablas y relaciones adecuadas para almacenar la información de manera eficiente y estructurada. Esto facilitará la posterior consulta y el análisis de los datos para extraer información relevante sobre el sector turístico.

- Análisis y explotación de los datos integrados mediante la construcción de consultas y/o visualizaciones analíticas: se explorará el desarrollo de *dashboards* interactivos que permitan visualizar de manera intuitiva y comprensible los aspectos más relevantes de la oferta vacacional en España. Estas visualizaciones facilitarán la toma de decisiones informada tanto para investigadores y planificadores turísticos como para los propios viajeros.

El logro de estos objetivos contribuirá al avance en la comprensión de las características y tendencias del sector turístico, así como a la mejora de la toma de decisiones en este ámbito.

1.3 Estructura de la memoria

A lo largo de la presente memoria se explica detalladamente el desarrollo del Trabajo de Fin de Grado. Este documento se estructura en los siguientes capítulos:

- **Capítulo 1:** Este primer capítulo es una introducción a lo que será el proyecto en sí. En él encontramos la motivación, los objetivos a cumplir y la estructura de la memoria.

- **Capítulo 2:** En esta sección de fuentes de datos y fundamentos se explican los diferentes aspectos teóricos para la correcta comprensión del proyecto. Así, en una primera parte se presentan las fuentes de datos utilizadas, para después pasar a introducir diferentes explicaciones en relación a los procesos ETL, los Data Warehouse y el concepto de análisis de datos, en una segunda parte.
- **Capítulo 3:** En este capítulo se incluyen los fundamentos tecnológicos del proyecto, en concreto, las diferentes herramientas y tecnologías utilizadas para llevarlo a cabo.
- **Capítulo 4:** Esta sección presenta las metodologías empleadas para desarrollar este trabajo de manera óptima y eficiente, así como la planificación hecha inicialmente y la finalmente realizada.
- **Capítulo 5:** Este apartado se centra en los procesos de análisis de requisitos de nuestro proyecto y de análisis de datos disponibles y su posterior selección.
- **Capítulo 6:** Esta sección recoge el proceso de creación del almacén de datos desde su diseño hasta su implementación, describiendo, además, los procesos ETL desarrollados.
- **Capítulo 7:** Este capítulo se dedica al análisis y explotación de datos mediante la creación de *dashboards* específicamente diseñados para extraer información de relevancia en el contexto de este proyecto. Además, se muestra la capacidad de poder realizar predicciones sobre ellos.
- **Capítulo 9:** Finalmente, se presentan las principales conclusiones de este trabajo, y posibles extensiones a futuro.

Capítulo 2

Fuentes de datos y fundamentos

Este capítulo presenta una introducción a diferentes aspectos teóricos del proyecto, desde las fuentes de datos utilizadas a la definición de diversos conceptos previos necesarios para su correcto entendimiento.

2.1 Fuentes de datos

Respecto a la selección de fuentes de datos se ha optado por Airbnb, como fuente de alojamientos, y el INE, como fuente de datos de turismo a nivel general.

La elección de Airbnb como fuente de datos viene dada por varias razones fundamentales. En primer lugar, esta plataforma ofrece un conjunto de datos accesibles al público. Además, respecto al sector turismo, Airbnb emerge como una elección natural, ya que es una de las plataformas más destacadas en este ámbito a nivel global y de las principales utilizadas a la hora de buscar alojamiento en nuestro país.

Por otro lado, en cuanto a datos globales de turismo, el Instituto Nacional de Estadística proporciona información de gran fiabilidad y facilidad de acceso.

2.1.1 Airbnb

Airbnb [1] es una de las plataformas de oferta vacacional más extendida en el mundo. Nació en 2008 y hoy día cuenta ya con millones de usuarios.

Por un lado están los anfitriones, los cuales ponen en oferta diferentes espacios, que pueden ir desde una casa rural hasta un ático de lujo. Por otro, los diferentes viajeros, los cuales utilizan la plataforma para realizar sus reservas sobre los alojamientos ofertados.

Airbnb ofrece, además, diferentes funciones [2] para que tanto el huésped como el anfitrión puedan tener una confianza plena en la reserva y se garantice una experiencia óptima:

- Verificación de la identidad del huésped.

- Análisis de las reservas: la plataforma analiza cientos de factores en cada reserva y bloquea aquellas que conllevan un riesgo elevado.
- Protección de 3 millones de dólares frente a daños: Airbnb te reembolsará por los daños que causen los huéspedes en tu alojamiento.
- Seguro de responsabilidad civil de 1 millón de dólares: cuentas con protección por si se da la situación de que un viajero tenga un problema, alguien le robe sus pertenencias o estas sufran algún desperfecto durante su estancia.
- Línea de protección 24 horas: equipo de ayuda por si surgiera cualquier problema.

El nivel al que se ha extendido esta plataforma en los últimos años y el crecimiento experimentado por la misma la convierten en un recurso de gran relevancia para el estudio y extracción de información en el ámbito turístico.

2.1.2 Instituto Nacional de Estadística

El Instituto Nacional de Estadística [3] es un organismo autónomo de carácter administrativo que tiene como objetivo proporcionar datos confiables y precisos que contribuyan a una toma de decisiones adecuada. Para ello produce información estadística de alta calidad en diversos ámbitos como pueden ser la economía, la sociedad, el medio ambiente o el turismo.

Esta información recolectada esta disponible en diversos formatos (tablas, series temporales, notas de prensa, ...) y tiene un acceso totalmente libre y gratuito a través de su web.

Debido a esta inmediatez que nos proporciona la web y a los diferentes datos turísticos que podemos encontrar en ella, constituye otra fuente de gran fiabilidad y calidad de datos para su uso y explotación combinada, junto con la anterior, en el desarrollo de este proyecto.

2.2 Conceptos previos

2.2.1 ¿Qué es el Web Scraping?

El Web Scraping es una técnica que tiene por objetivo extraer información de uno o varios sitios web y procesarla en un formato más simple y comprensible, como pueden ser hojas de cálculo o archivos CSV.

La característica principal de este proceso es que simula la navegación web humana para recopilar dichos datos e información y su ventaja se encuentra en su capacidad para ser programado y/o automatizado.

Esta técnica se puede implementar con diferentes lenguajes de programación. Para la realización de este proyecto se utilizará Python con el apoyo de la biblioteca BeautifulSoup, muy

destacada en este campo, ya que permite analizar el código HTML de las páginas web, identificar los elementos de interés y extraer la información relevante.

2.2.2 ¿Qué es un proceso ETL?

Generalmente una base de datos obtiene sus datos de diferentes fuentes, siendo así muy probable que tengan también diferente naturaleza y no se pueda realizar su integración directamente. En estas situaciones es necesario realizar un proceso ETL [4, 5] (Figura 2.1) para llevar a cabo la integración de los datos y garantizar la calidad de los mismos.

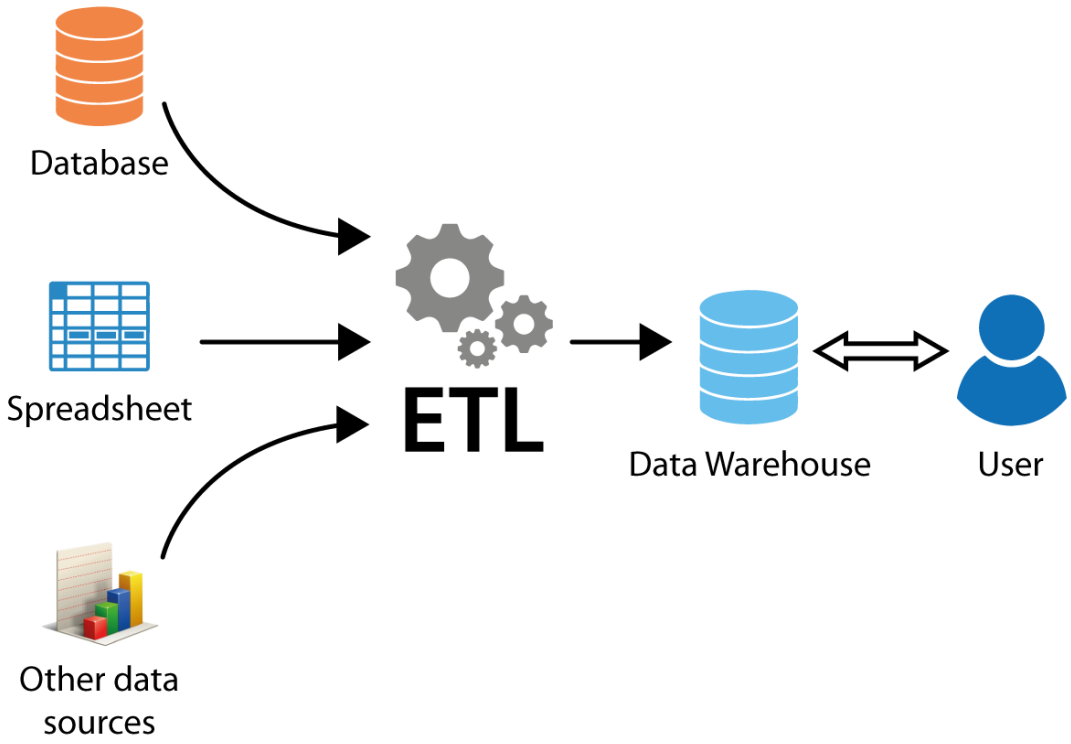

Figura 2.1: Proceso ETL

Fases ETL

El proceso consta de tres etapas (Figura 2.2):

- Extracción: En esta fase se lleva a cabo la extracción de los datos necesarios de las diferentes fuentes que se hayan seleccionado.
- Transformación: Tras realizar la extracción se procede a hacer un primer análisis para llevar a cabo una limpieza y corrección de los datos, pudiendo eliminar registros duplicados, normalizar datos o convertir formatos, entre otras transformaciones. Así se podrá posteriormente realizar un estudio de mayor eficiencia y calidad.
- Carga: La última etapa se basa en cargar los datos en el destino final, dónde estarán disponibles para su posterior explotación.

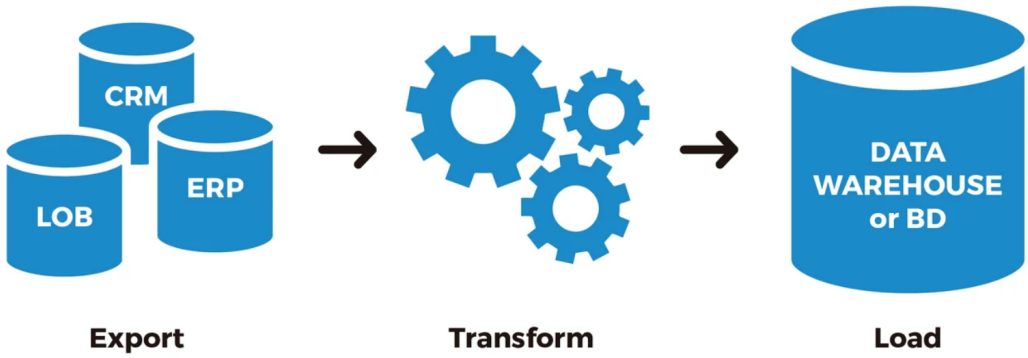

Figura 2.2: Fases ETL

2.2.3 ¿Qué es un Data Warehouse?

Un Data Warehouse [6] (Almacén de datos, Figura 2.3) es una infraestructura de gestión y almacenamiento de datos diseñada especialmente para soportar tareas analíticas en grandes volúmenes de información. Además, los datos generalmente provienen de diversas fuentes; el almacén de datos optimizará la integración de los mismos para luego poder explotarlos.

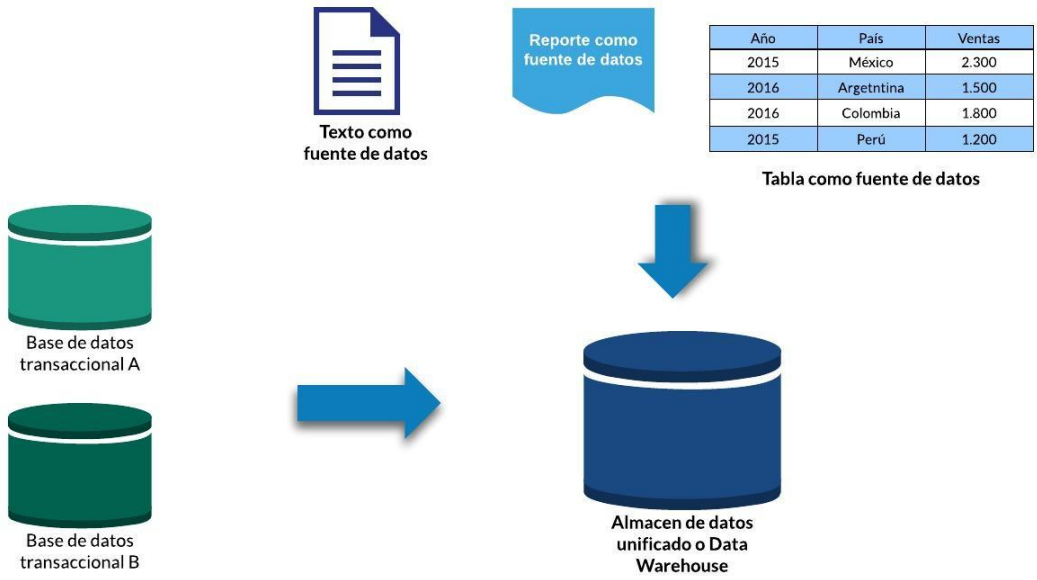

Figura 2.3: Data Warehouse

Estructura

Un esquema muy recurrido para la representación de un Data Warehouse es el modelo relacional, dicho modelo está compuesto por tablas que pueden ser de dos tipos:

- Tablas de dimensiones: en estas tablas encontramos representados los factores mediante los que se llevará a cabo el análisis de los diferentes objetivos o áreas de negocio. A continuación, veremos en nuestro proyecto diferentes tablas de dimensiones, como puede ser la de alojamientos, en ella podremos encontrar su evaluación, capacidad, ...
- Tablas de hechos: éstas están relacionadas con las anteriores a través de campos clave y en ellas encontramos aquéllo que se va a analizar. En nuestro caso encontraremos como medida, por ejemplo, el precio por noche.

En cuanto a los tipos de modelos que se pueden formar destacan tres [7]:

- Esquema de estrella: Es el modelo más simple y su diagrama, como bien podríamos intuir, representa una estrella. El centro de éste contiene una tabla de hechos, la cual se relacionará con diferentes tablas de dimensiones. Este esquema tiene poca complejidad de consulta pero tendrá múltiples relaciones de datos.
- Esquema de copo de nieve: El concepto de jerarquía se introduce en este modelo ya que la tabla de hechos se conectará a diferentes tablas de dimensiones, las que a su vez se conectarán con otras tablas de dimensiones, 'subdimensionales'. Los datos estarán normalizados y no serán redundantes ni habrá repeticiones.

- Esquema de galaxia: Este esquema es más complejo. En él varias tablas de hechos comparten dimensiones; las tablas de hechos no necesitan estar directamente relacionadas. Da lugar a un esquema flexible de alta calidad y precisión de datos y podríamos decir que es un esquema híbrido de los dos anteriores, ya que combina elementos de estos.

2.2.4 Análisis de datos

El análisis de datos [8] es el proceso mediante el cuál queremos obtener información útil, monitorear, predecir y poder respaldar la toma de decisiones con el fin de obtener beneficios en base a ello. Este proceso implica aplicar una variedad de técnicas y herramientas para extraer significado de los datos e identificar patrones, tendencias o relaciones.

El análisis de datos se puede dividir en varios tipos [9] en base a sus diferentes propósitos; en este proyecto se trabajará con el análisis descriptivo, su objetivo final es describir un conjunto de datos. Mediante éste podremos obtener parámetros que distinguen las características del conjunto de datos y conocer detalladamente la información que poseemos.

Capítulo 3

Herramientas y tecnologías utilizadas

En este capitulo se describen las diferentes herramientas y tecnologías más destacables y necesarias para el correcto desarrollo del proyecto.

3.1 Herramientas

3.1.1 Visual Studio

Visual Studio [10] es una plataforma que tiene como principal objetivo editar, depurar y compilar código. Además acepta múltiples lenguajes de programación e incluye multiples funciones para mejorar el proceso, como pueden ser el completado de código o diseñadores gráficos.

Esta herramienta ofrece soporte tanto para Windows como para MacOS, Linux o Web y ha sido desarrollada por Microsoft.

3.1.2 Talend

Talend [11] es un software comercial de integración de datos, de código abierto y bajo licencia. Esta solución surge motivada por la gran cantidad de datos masivos que forman parte de cualquier organización, con objetivo de medir la salud de los mismos y de desarrollar una mejor gestión, y mayor confianza en ellos.

Talend combina la integración, la integridad y la gobernanza de los datos en una única plataforma unificada que ofrece todo el valor de los datos para su uso y explotación. Entre sus características cabe destacar:

- Ahorra recursos acelerando los datos operativos.

- Moderniza la nube y los datos para que los diferentes negocios puedan crecer.
- Reduce el riesgo y establece la excelencia de los datos.

3.1.3 PowerBI

Power BI [12] es una plataforma unificada y escalable de inteligencia empresarial (BI) con principal objetivo la generación de informes. Permite tanto unir diferentes fuentes de datos y modelarlas como llevar a cabo su posterior análisis y crear paneles de visualizaciones para extraer conclusiones de manera intuitiva.

Esta herramienta cuenta con multitud de aplicaciones y servicios basados en la nube, además de la más conocida aplicación de escritorio, PowerBI Desktop.

3.1.4 R Studio

R Studio [13, 14] es un entorno gráfico que simplifica la creación y ejecución de scripts para el lenguaje de programación R.

R Studio utiliza una partición de la pantalla en diferentes secciones facilitando asi la visualización de código fuente, datos, resultados obtenidos, gráficos, etc. También facilita la creación de informes en diferentes formatos (principalmente, HTML o PDF)

3.2 Tecnologías

3.2.1 MySQL

MySQL [11] es el sistema de administración de bases de datos SQL, de código abierto, más popular y es Oracle Corporation quién lo ha desarrollado, distribuido y respaldado. Las bases de datos MySQL son relacionales, es decir, almacenan datos en tablas separadas, en vez de poner todos los datos en un gran almacén.

Respecto a SQL destacar que es un lenguaje de consulta estructurado y el más utilizado para acceder a bases de datos.

3.2.2 Python

Python [15, 16] es un lenguaje de programación de libre uso y distribución ya que se ha desarrollado bajo una licencia de código abierto administrada por Python Software Foundation.

Este lenguaje es utilizado en diversos campos como son las aplicaciones web, el desarrollo de software, la ciencia de datos y el aprendizaje máquina. Su uso es ampliamente extendido porque se trata de un lenguaje que destaca debido a su eficiencia y facilidad de aprendizaje.

Python cuenta además con librerías reutilizables para diversas tareas que se pueden encontrar en una gran biblioteca estándar. En el desarrollo de este proyecto cabe destacar dos:

- Beautiful Soup [17]: su principal cometido es extraer datos de archivos HTML y XML. Dado que en el proyecto se necesita extraer información de las páginas web, necesitaremos una librería que acceda a los datos de los archivos HTML.
- Selenium [18]: esta librería permite interactuar con diferentes navegadores a través de un controlador, un Web Driver.

3.2.3 Pandas

Pandas [19] es un objeto Data Frame rápido y eficiente que ofrece herramientas para la manipulación de datos. Con Pandas podremos, entre otras funcionalidades:

- Leer y escribir datos
- Gestionar datos faltantes
- Transformar datos
- Segmentar en función de etiquetas
- Fusionar y unir conjuntos de datos
- Usar de manera eficiente series temporales

En conclusión, Pandas busca ser la herramienta fundamental de código abierto para el tratamiento y análisis de datos en Python, por ello será de gran interés en el desarrollo de este proyecto.

GeoPandas

El proyecto GeoPandas [20], también de código abierto, busca dar soporte a datos geográficos de los objetos Pandas. Éste puede tratar con las subclases de Pandas 'pandas.Series' y 'pandas.DataFrame' implementando GeoSeries y GeoDataFrame, respectivamente.

En este proyecto esta herramienta será de ayuda para, como se verá más adelante, poder ubicar cada uno de los alojamientos.

Capítulo 4

Metodología y Planificación

4.1 Metodología

El desarrollo de este trabajo de fin de grado sigue, principalmente, una metodología iterativa.

4.1.1 Iterativa

La metodología iterativa consiste en dividir el proyecto en diferentes bloques temporales llamados iteraciones; de esta manera, vamos completando los objetivos de forma incremental en cada iteración.

Para su correcto desarrollo y funcionamiento se deben establecer los diferentes objetivos en base a las fases del proyecto, y otorgarles un orden según las necesidades a satisfacer. Por tanto, de acuerdo con ello, inicialmente, se han determinado los distintos objetivos y necesidades de análisis para, posteriormente, poder proceder con el desarrollo de las diferentes fases del proyecto que veremos a continuación.

La decisión de utilizar esta metodología para el desarrollo de este proyecto atiende a diferentes motivos/beneficios:

- Se pueden gestionar los cambios de manera natural. Es decir, al finalizar cada iteración se obtiene una visión del proyecto desde la cual es posible observar como éste va avanzando y gestionar los cambios necesarios de cara al futuro para un correcto desarrollo del trabajo.

- Permite llevar a cabo un seguimiento sencillo del progreso desde las primeras iteraciones, pudiendo así comprobar si su finalización en la fecha prevista es viable y adaptar los tiempos de cada iteración según vayan surgiendo imprevistos.

- Permite mitigar riesgos desde el principio. Una consecuencia del punto anterior es ésta;

llevar a cabo un seguimiento desde el inicio nos permite anticiparnos y mitigar riesgos antes de que se hagan demasiado relevantes.

- Podemos gestionar la complejidad del proyecto, ya que la dividimos en iteraciones.

4.1.2 Diseño de las iteraciones

Para llevar a cabo la adaptación de la metodología antes mencionada al proyecto, se ha optado por dividir éste en fases, definiendo así las diferentes iteraciones del modelo.

- **Iteración 1: Análisis de requisitos y estudio y selección de datos.** En esta fase se hace un estudio inicial y se seleccionan las variables y características más relevantes para extraer y utilizar posteriormente en el análisis de los datos.
- **Iteración 2: Extracción de datos.** En esta fase se realiza la extracción de datos de las fuentes de información. Como se verá más adelante, durante la ejecución de este proyecto se decidió llevar a cabo esta iteración en dos partes, realizando primeramente la extracción de un pequeño conjunto de datos con el que poder ir programando el proceso ETL, al tiempo que después se obtenía el conjunto de datos completo.
- **Iteración 3: Diseño e implementación de procesos para la transformación y limpieza de datos.** Se diseñan y realizan tareas de limpieza y transformación de los datos para asegurar mayor calidad y precisión, ya que los datos recopilados suelen contener errores o información redundante. También se transforma el formato de los datos para poder analizarlos de manera más eficiente y poder integrarlos correctamente.
- **Iteración 4: Diseño e implementación de un almacén para la integración de los datos.** Se crea un modelo relacional óptimo para llevar a cabo el proceso de carga e integración de los datos de manera eficiente.
- **Iteración 5: Transformación y carga.** Una vez diseñados los diferentes procesos de transformación y carga de los datos, se procede a realizar los mismos.
- **Iteración 6: Análisis y visualización de datos.** En este punto se obtiene información valiosa del conjunto de datos, pudiendo visualizar la información obtenida a través de gráficos y tablas para facilitar su comprensión.
- **Iteración 7: Extracción de conclusiones.** Una vez realizado el análisis y visualización de datos, se extraen conclusiones sobre los patrones y tendencias observadas. Se identifican las variables más importantes y se elaboran hipótesis sobre los factores que pueden estar afectando a los resultados.

- **Iteración 8: Elaboración de la memoria.** Todo el proceso se documenta, incluyendo información sobre objetivos, métodos utilizados, resultados obtenidos, conclusiones extraídas y cualquier otra información de relevancia. Esta iteración, realmente, se lleva a cabo desde el inicio hasta el final del proyecto, a fin de facilitar la correcta y completa documentación del mismo.

4.2 Planificación

Para llevar a cabo correctamente un proyecto de estas características, con importante carga de trabajo, debe realizarse una planificación inicial para poder abordar de manera óptima el proceso completo.

Así, se ha creado un diagrama de Gantt inicial donde poder reflejar las diferentes tareas de una manera más visual, abarcando toda la vida del proyecto. Cada tarea define un inicio y un fin estimados en base a su dificultad o el tiempo mínimo requerido para su realización. Cabe destacar que las diferentes tareas son evoluciones del sistema y no etapas estancas, independientes las unas de las otras.

El diagrama de Gantt correspondiente a la planificación inicial de este proyecto puede observarse en la Figura 4.1

Figura 4.1: Diagrama Gantt planificación inicial

Aunque es importante intentar cumplir la planificación, también es natural que, en proyectos de estas dimensiones, puedan surgir imprevistos que provoquen desviaciones en los tiempos de realización de una tarea, pudiendo tanto retrasarlas como adelantarlas, ya que estos tiempos son simples estimaciones.

En el diagrama que se muestra en la Figura 4.2, correspondiente a la planificación final de este proyecto, podemos apreciar algunas de esas desviaciones. Destaca entre ellas la demora experimentada en el proceso de extracción de datos, que a su vez generó un efecto en cadena. Aún así, como ya se adelantó anteriormente, fue posible seguir avanzando en otras tareas

mientras concluía la extracción del conjunto de datos global, realizando las diferentes tareas con un pequeño conjunto de datos de ejemplo representativos de todas las características. Con él se pudo ir programando el proceso de limpieza y realizar el diseño del almacén. Para reflejar esta circunstancia, en el diagrama final se muestra la división del proceso general de extracción en una primera extracción inicial y una extracción final.

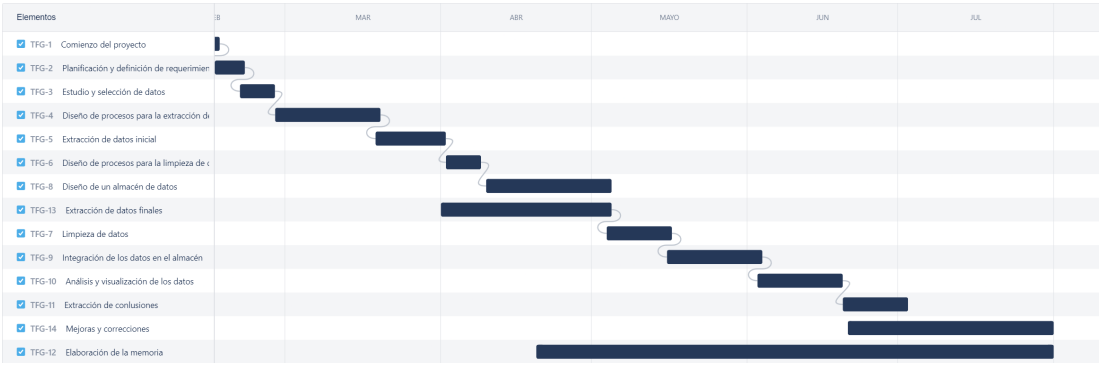

Figura 4.2: Diagrama Gantt planificación final

Se añade también a cambios contemplados sobre la planificación inicial, la implementación de diferentes mejoras, surgidas durante la realización del proyecto, relativas, entre otras, al proceso de incorporación de nuevo datos tras tu extracción, y que supusieron su extensión en el tiempo.

Costes

En cuanto al tema de costes del proyecto, por un lado tenemos los recursos técnicos empleados en material. Incluimos un ordenador propio, del que ya se disponía, y las herramientas mencionadas en el Capítulo 3, que tampoco suponen ningún tipo de coste, ya que se trata de soluciones abiertas.

Respecto a las horas trabajadas diariamente, éstas no han sido constantes. En un inicio, el primer mes se han invertido en el proyecto únicamente dos horas diarias de lunes a viernes; a partir de este momento, concretamente a partir del día 20 de marzo, este número ha aumentado a 3 horas diarias y sumando un día más a la semana. Finalmente, aunque inicialmente la realización de este proyecto debería llevar alrededor de 300h, se aproxima a un total de 330h. Tras esta recapitulación de horas invertidas y estimando un coste de 25 €/h se trataría de un proyecto que ronda un coste de 8250€ para el autor, el científico de datos.

Finalmente, hay que añadirle el coste estimado de los directores de proyecto, derivado de las horas empleadas en reuniones para la revisión del mismo y la resolución de dudas. El precio por hora estimado para un director rondaría los 35€/h. Sumando un total de alrededor de 30 horas por cada uno de los directores, quedaría en 2100€ a mayores, lo que resultaría en

un coste global de 10350€. Podemos ver este coste en la Tabla 4.1.

Recursos	Costes iniciales	Costes finales
Directores de proyecto	2100 €	2100 €
Científico de datos	7500 €	8250 €
Herramientas y tecnologías	0 €	0 €
TOTAL	**9600 €**	**10350 €**

Tabla 4.1: Costes estimados

Capítulo 5

Análisis

En este capítulo se tratan los primeros pasos del proceso de análisis de datos; es decir, se aborda el problema de la obtención de datos. Así, tras un análisis de requisitos inicial, se determinan y extraen los datos adecuados para llevar a cabo el posterior proceso de análisis y se procede a su limpieza.

5.1 Análisis de requisitos

Lo más importante antes de iniciar un proyecto es determinar los requisitos y las características de mayor relevancia del mismo. A día de hoy contamos con una inmensa cantidad de datos de diferentes fuentes y muy variados, por tanto hacer una selección de los aspectos clave es fundamental para la realización de un buen análisis posterior.

Con este proyecto se busca conocer la oferta turística vacacional que nos proporciona hoy en día la creciente aplicación de Airbnb y estudiar, además, posibles relaciones con otros datos del sector procedentes del INE. La utilización de ambas fuentes de información es complementaria y permite abordar distintos aspectos de interés en torno al ámbito turístico.

5.1.1 Necesidades de análisis

En este proyecto se han observado diferentes puntos relevantes de análisis a tener en cuenta:

- **Comparativas entre INE y Airbnb.** Dado que el INE nos brindará información acerca del sector turismo del país y Airbnb nos aportará una visión de la oferta de alojamientos turísticos, como es lógico, es de interés encontrar correlaciones entre los datos de ambas fuentes. Dichas correlaciones pueden ser el número de registros de nuevos alojamientos según los gastos en turismo, el número de turistas y capacidades de los alojamientos, gastos y duraciones de viajes o precios por noche, entre otros. Como bien se menciona en la Sección 1.1 de motivación, esto servirá, además, para complementar las encuestas

y datos actuales recolectados por el INE con diferentes características sobre los alojamientos ofertados hoy en día en nuestro país.

- **Gastos y precios.** A la hora de la toma de decisiones en la mayoría de contextos podemos encontrarnos la importancia que tiene el dinero y cómo mueve muchas de estas decisiones, por tanto, otro punto a tratar será el de los gastos en turismo y precios de alojamientos. Conocer el precio medio y cómo se relaciona con los servicios podrá permitir a los diferentes propietarios establecer precios competitivos y que se adapten a las capacidades. A mayores, este análisis de gastos en turismo y precios podrá revelar oportunidades de negocio, por ejemplo, lugares en los que los precios de alojamientos sean más altos los emprendedores pueden desarrollar modelos de negocio innovadores, como alojamientos compartidos o servicios de bajo costo, para aprovechar esta oportunidad y atraer a un nuevo grupo de clientes.

- **Galicia.** Dado el entorno geográfico en que nos ubicamos, se ha incluído como elemento de interés el análisis de información de Galicia en comparación a España, para poder ver algunas diferencias que pueda tener nuestra comunidad con el resto del país. Este análisis podrá ayudar en la planificación y desarrollo turístico a nivel regional, conociendo diferencias relevantes con respecto al resto de la nación que podrán ayudarnos a crecer en este sector.

- **Repercusión de las evaluaciones.** Finalmente, centrándonos en la plataforma seleccionada para el estudio, Airbnb, también es un dato de relevancia la importancia que la aplicación le da a los usuarios viendo cómo influyen las puntuaciones que estos aportan a los diferentes alojamientos. Al propietario del alojamiento le brindará ayuda para la toma de decisiones, ya que le aportará información valiosa acerca de si le compensa implementar ciertos servicios o no, si tener buenas evaluaciones le permitirá incrementar el precio del alojamiento o incluso poder alcanzar la insignia que Airbnb ofrece de Super Anfitrión.

5.2 Análisis de los datos

El objetivo de esta sección es analizar las características útiles de estudio para este proyecto.

5.2.1 Airbnb

Dados los supuestos de estudio y necesidades de análisis previamente planteadas y analizados los elementos informativos de los alojamientos disponibles a través de la web de Airbnb, se ha determinado extraer el siguiente conjunto de características:

Características de los alojamientos

A continuación se enumeran las características relacionadas con los propios alojamientos:

- Nombre: esta característica quizás no es la más relevante como valor a analizar; sin embargo, se ha decidido incluír debido a su capacidad para distinguir inicialmente los alojamientos y tener referencias de cada uno de ellos.
- Capacidad máxima: esta característica brindará información para poder evaluar el número de viajeros que cabría esperar y ver la influencia que ello puede tener.
- Número de baños, camas y dormitorios: estos tres valores aportarán información de las dimensiones y capacidades del alojamiento, las cuales se podrán utilizar para estimar cómo afectan a otros campos.
- Fecha de registro del alojamiento: saber el volumen de registros a lo largo de los años y conocer la evolución de los mismos se ha presentado como un análisis de gran utilidad.
- Latitud y longitud (aproximadas): Airbnb no ofrece la posibilidad de conocer la ubicación exacta de un alojamiento hasta que se realiza la reserva; no obstante, sí que da la opción de ver unas coordenadas aproximadas para conocer el lugar o la zona en que se localiza.
- Servicios: aquí se almacenan los diferentes servicios que ofrecen cada uno de los apartamentos. Entre ellos se encuentran la disponibilidad de wifi, aparcamiento o calefacción, los cuales previsiblemente tengan influencia en el precio y las evaluaciones del alojamiento.

Características del alquiler

En esta sección se lista la información vinculada al propio alquiler (el precio del mismo, la satisfacción tras la reserva, etc.).

- Precio por noche: conocer el precio de los alojamientos, como es lógico, es un campo muy relevante, ya que estará relacionado con muchas otras características, como son los servicios, el tipo de anfitrión, las evaluaciones, la ubicación, etc. Además, éste es el principal factor para la elaboración de estrategias de negocio, establecimiento de precios competitivos y toma de decisiones futuras.
- Tiempo y ratio de respuesta: un valor de interés para el usuario es conocer la velocidad de respuesta de un propietario y si éste acostumbra a darla, puesto que, en caso contrario, quizás se tome la decisión de seleccionar otro alojamiento en el que el anfitrión esté más pendiente de sus huéspedes.

- Evaluación media: esta característica es de gran valor, ya que es un indicador de cómo de satisfecho se queda un usuario tras abandonar el alojamiento.
- Número de evaluaciones: dato relevante para el análisis de evaluaciones, ya que no es lo mismo un alojamiento que cuente con una única evaluación de 5 estrellas que otro que tenga una media de 4,6 estrellas, pero obtenida a partir de 1000 evaluaciones. En este contexto, por ejemplo, dará más seguridad el segundo alojamiento.

Otras características de interés

Seguidamente se describen otras características a mayores que serán interesantes para llevar a cabo un análisis más profundo.

- Tipo de cancelación, hora de llegada y salida, idiomas: se ha optado por almacenar también estas tres características debido a que pueden ser de relevancia en un trabajo futuro más exhaustivo, centrándose principalmente en cada uno de los alojamientos, su rendimiento y propiedades.
- Tipo de anfitrión: así como el nombre del anfitrión no aporta gran valor y se decide no incluirlo, saber si es Super Anfitrión o no, servirá para hacer comprobaciones acerca de su influencia en otros valores como pueden ser las evaluaciones o el precio.
- Fecha de extracción: a mayores de los datos ofrecidos por la propia aplicación hemos visto importante etiquetar el mes y año de extracción de cada alojamiento para así poder ver la evolución de la plataforma a lo largo del tiempo con extracciones futuras.

Además, se ha optado por descartar ciertas características. De entrada, el nombre detallado y la descripción del alojamiento no se consideraron ya que no nos iban a aportar valor suficiente, ni ahora ni en futuras extensiones a este trabajo, en comparación con el peso de almacenar dichos datos. La descripción tiene un tamaño considerable. Por otro lado, por los motivos mencionados anteriormente, también se descartó el nombre del anfitrión; lo mismo que las características relativas al hecho de calificar la cama donde se duerme como XL, XXL o simplemente grande, ya que se ha considerado un valor relativo. Los comentarios tampoco han sido almacenados debido a su gran volumen en comparación con el valor que se consideró que aportaría, al igual que la disponibilidad del alojamiento, en este caso, debido también a que su extracción daría lugar a posibles dificultades y es un valor demasiado variable. En el momento de extracción, un alojamiento puede estar libre en muchas fechas y, a la hora siguiente, por la publicación de un anuncio, podría ser reservado por diferentes usuarios en diversos periodos.

Por último, en cuanto a la ubicación, se ha descartado recoger la localización como característica, ya que en dicho campo cada propietario indica la que él considera y no llega a estar

completamente normalizada. Sí se decidió, sin embargo, extraer las coordenadas aproximadas (latitud y longitud) del alojamiento, pues resultan relevantes para el análisis, al permitir vincular cada alojamiento a un municipio y hacer un estudio por localizaciones.

5.2.2 INE

En cuanto a los datos del propio sector turismo del país, se ha obtenido la información del Instituto Nacional de Estadística. Tras realizar un análisis de la web y comprobar la disponibilidad y oferta de datos, se decidió complementar los datos de Airbnb con los que se pueden ver en la Tabla 5.1, extraídos del Instituto Nacional de Estadística [21, 22]. Para cada Comunidad Autónoma se disponía de la información, por años, mostrada en la tabla.

Características
Comunidad Autónoma de destino
Año
Duración media del viaje y su tasa de variación anual
Gasto medio por persona y su tasa de variación anual
Gasto medio diario por persona y su tasa de variación anual
Gasto total y su tasa de variación anual
Número de turistas y su tasa de variación anual

Tabla 5.1: Características turismo INE

Teniendo todos estos datos en cuenta se decidió eliminar las tasas de variación anual, ya que se podrían obtener con un simple cálculo entre un valor y su correspondiente en el año anterior, quedándonos así con:

- Duración media del viaje: dato relevante para conocer si el precio y gasto en un lugar interfiere en la duración de los viajes.
- Gasto medio por persona, gasto medio diario por persona y gasto total: todos estos gastos serán relevantes como medida para, al igual que sucedía con el precio por noche, buscar las Comunidades Autónomas más caras y poder aportar alternativas turísticas más económicas; ver también si hay relación con las variables de duración y, de ser así, analizar si abaratando los gastos aumentan las duraciones, etc.
- Número de turistas: este dato es de interés a la hora de comprobar capacidades de alojamientos o regiones con más turismo. En el caso de las comunidades con escaso turismo, trabajadores de este sector podrían hacer un estudio de mercado de las comunidades con mayor turismo y analizar sus flaquezas de las que tienen poco.
- Comunidad Autónoma de destino y año: necesarios para organizar los datos y hacer comparaciones entre regiones o ver su evolución en el tiempo.

Igualmente, se decidió añadir también a este conjunto de características el gasto total mensual en España [23] correspondiente al sector turismo, aunque dicha información no estuviese disponible a nivel de Comunidades Autónomas, para así analizar la evolución no solo a lo largo de los años sino también a nivel de mes.

Además de las características señaladas, el INE contaba con otros datos que se han decidido descartar, como pueden ser el número de turistas por tipo de viaje o los índices de ocupación y precios de los diferentes tipos de alojamientos. Esta información no fue considerada por diversos motivos. En el caso del número de turistas por el tipo de viaje, no se consideró lo suficientemente relevante para el análisis, ya que el objetivo no es tratar los diferentes tipos de viaje. Por otro lado, los datos referentes a los tipos de alojamiento no se tienen en cuenta, ya que en la plataforma de Airbnb no se dispone de esta distinción y, aunque la hubiese, un anfitrión podría calificar su alojamiento como considerase, subjetivamente.

Todas las características señaladas constituyen el punto de partida para la posterior realización de estudios con los que abordar los diferentes análisis exploratorios, pudiendo observar cómo afecta el precio y el gasto a las diferentes decisiones de los viajeros, comparar las diferentes fechas y comunidades a la hora de viajar o analizar cómo las distintas evaluaciones y propiedades de los alojamientos tienen también efecto sobre ello.

Capítulo 6

Diseño e implementación del almacén de datos

6.1 Diseño del modelo conceptual

Una vez determinadas las características necesarias y relevantes para alcanzar los objetivos planteados se procede a realizar un modelo conceptual del que será el almacén de datos. Dicho modelo servirá para obtener un esquema detallado de lo que será el Data Warehouse, ordenándolo en hechos y dimensiones. Esta descripción del almacén dará información tanto sobre los datos que se recogen en él como de las diferentes relaciones entre ellos.

En la Figura 6.1 se puede ver el esquema elaborado para estructurar el Data Warehouse de este proyecto en el que se encuentran las siguientes tablas de hechos y dimensiones.

Hecho alquiler. Hecho que recoge las métricas para poder hacer el estudio de los alquileres en el país y se relaciona con las dimensiones de alojamientos, anfitriones, cancelación y fecha.

- **id_alojamiento:** Clave foránea que nos permitirá obtener la información necesaria del alojamiento relacionado con cada alquiler.
- **id_anfitrion:** Clave foránea vinculada a los datos del anfitrión.
- **id_cancelacion:** Clave foránea relativa a la cancelación de la reserva.
- **id_fecha:** Clave foránea que nos dará a conocer la fecha en la que se han recogido los datos.
- **price_per_night:** Medida que indica el precio por noche del alquiler.
- **n_evals:** Número de evaluaciones del apartamento.
- **media_ratings:** Número medio de estrellas obtenidas; tiene un valor de entre 0 y 5.

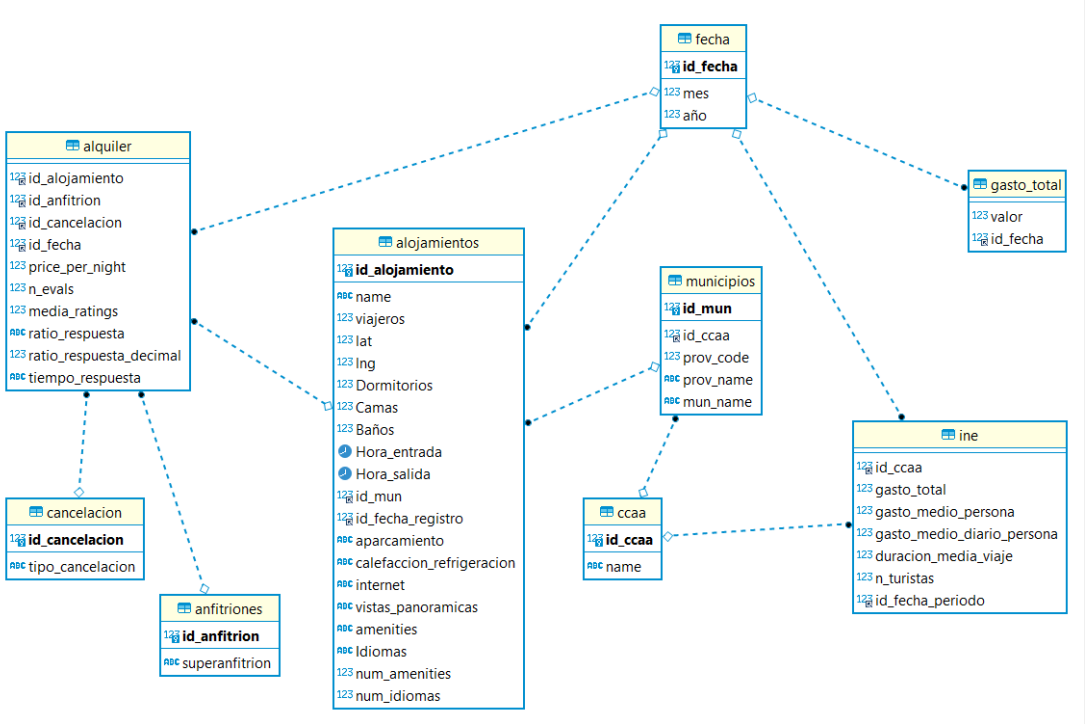

Figura 6.1: Modelo conceptual Data Warehouse

- **Ratio de respuesta:** Porcentaje de probabilidad de respuesta que tendrán los usuarios al ponerse en contacto con el alojamiento.
- **Tiempo de respuesta:** El tiempo que tardarán los usuarios en obtener respuesta del mismo.

Hecho ine. En este hecho se encuentran las métricas necesarias para realizar un estudio de los datos de turismo, en las diferentes comunidades de España, obtenidas a través del INE.

- **id_ccaa:** Clave foránea que nos permite relacionar esta tabla con la de Comunidades Autónomas y así obtener las medidas de cada una de ellas.
- **gasto_total:** Gasto total de los turistas en las diferentes comunidades de destino, en millones de euros.
- **gasto_medio_persona:** Gasto medio por persona en el destino, en euros.
- **gasto_medio_diario_persona:** Gasto medio diario por persona en el destino, en euros.
- **duracion_media_viaje:** Duración media de los viajes.
- **n_turistas:** Número de turistas que han ido a dicho destino.

- **id_fecha_periodo:** Año al cual corresponden las diferentes métricas de turismo; estará asociado al mes de Diciembre del año correspondiente.

Hecho gasto_total. En este hecho se incluye la medida correspondiente al gasto total por meses, en vez de por Comunidades Autónomas, ya que no había posibilidad de obtener ambas conjuntamente a través del INE.

- **id_fecha:** Clave foránea que nos permite relacionar esta tabla con la de fechas y nos indica a qué mes y año corresponde ese gasto.
- **valor:** Gasto total de los turistas por meses, en millones de euros.

Dimensión de localización. Dimensión que, cómo bien indica su nombre, recoge la información relacionada con las diferentes localizaciones de España. Esta dimensión trabaja en dos granularidades y por tanto se divide en dos tablas, una a nivel de comunidades autónomas y otra a nivel de municipios. En la primera de ellas se encuentran:

- **id_ccaa:** Código único asignado a cada una de las comunidades, clave primaria.
- **name:** Los diferentes nombres de las comunidades autónomas.

A continuación, se listan las características que recoge la tabla de los diferentes municipios de España.

- **id_mun:** Identificador único de cada municipio, clave primaria.
- **mun_name:** Nombre del municipio
- **prov_code:** Código de la provincia a la que pertenece dicho municipio.
- **prov_name:** Nombre de la provincia a la que pertenece dicho municipio.
- **id_ccaa:** Clave foránea que relaciona esta dimensión con la de Comunidades Autónomas pudiendo así identificar en qué Comunidad Autónoma se encuentra cada municipio.

Dimensión de fecha. En esta dimensión se recoge la fecha de diferentes características.

- **id_fecha:** Identificador único asociado a cada fecha, clave primaria.
- **mes:** Mes del año.
- **año:** Año.

Dimensión de alojamientos. Esta dimensión tan amplia recoge información de los diferentes alojamientos de España; en ella se ven también la mayoría de características de los mismos referidas en el Capítulo 5.

- **id_alojamiento:** Clave primaria y, por tanto, identificador único de cada alojamiento.
- **name:** Nombre de cada alojamiento, dada por el anfitrión.
- **viajeros:** Número de viajeros máximos; es decir, la capacidad del alojamiento.
- **amenities:** Diferentes características extra asociadas al alojamiento.
- **lat:** Latitud.
- **lng:** Longitud.
- **Dormitorios:** Número de dormitorios.
- **Camas:** Número de camas.
- **Baños:** Número de baños.
- **Idiomas:** Idiomas admitidor por el anfitrión para establecer una comunicación con el/ella.
- **Hora_entrada:** Hora de llegada al alojamiento establecida.
- **Hora_salida:** Hora a la que se debe haber abandonado la estancia el último día de reserva.
- **num_amenities:** Número de características extras con las que cuenta el alojamiento.
- **num_idiomas:** Número de idiomas con los que es posible comunicarse a la hora de la reserva.
- **internet:** Nos informa de si el alojamiento cuenta con conexión a Internet.
- **vistas_panoramicas:** Indica si el alojamiento tiene vistas panorámicas.
- **calefaccion_refrigeracion:** Nos dice si el alojamiento cuenta con servicio de calefacción y aire acondicionado.
- **aparcamiento:** Nos informará si podremos disponer de aparcamiento durante la estancia.
- **id_mun:** Clave foránea a la tabla de municipios, la cual nos permitirá establecer el municipio concreto al que pertenece cada alojamiento.

- **id_fecha_registro:** Fecha en la que se ha registrado el alojamiento en la plataforma.

Entre las características enumeradas nos encontramos con medidas que se podrían encontrar en una tabla de hechos, como pueden ser el número de evaluaciones o de amenities. En aras de simplificar el modelo, se ha decidido introducir todas estas características en la dimensión de alojamiento, en vez de crear nuevas tablas de hechos para ellas.

Dimensión de cancelación. En esta dimensión se recogen los tipos de cancelación que puede sufrir un alquiler.

- **id_cancelacion:** Identificador único para cada tipo, clave primaria.
- **tipo_cancelacion:** Tipo de cancelación; puede ser reembolso parcial, total o no aceptar reembolso.

Dimensión de anfitriones. Esta tabla contiene la información referente a los anfitriones.

- **id_anfitrion:** Clave primaria e identificador único de cada tipo de anfitrión.
- **superanfitrion:** Los anfitriones pueden ser Super Anfitriones o no, y ese dato se recoge aquí.

6.2 Proceso ETL

Una vez se tienen claras las características y se ha diseñado un modelo lógico que dará forma al almacén para alcanzar los objetivos planteados, se procede a realizar el proceso ETL enfocado en la consecución de esos objetivos.

Como se ha mencionado en la Sección 2.2.2, este proceso se divide en tres fases (Figura 6.2). En la primera fase de extracción se obtienen los datos necesarios con el apoyo de técnicas de *Web scraping*. A continuación, se procede con la adaptación de los datos para poder trabajar con ellos, su transformación y limpieza. Finalmente, se lleva a cabo la carga de los mismos en el almacén.

6.2.1 Extracción de características

Para extraer el volumen de datos que conforma el núcleo del proyecto a nivel de información se han llevado a cabo dos procesos, uno para cada plataforma objeto de tratamiento.

Extracción de datos del INE

El proceso de extracción de los datos del INE resulta inmediato gracias a que esta plataforma proporciona acceso a los datos en diversos formatos de interés; entre ellos archivos

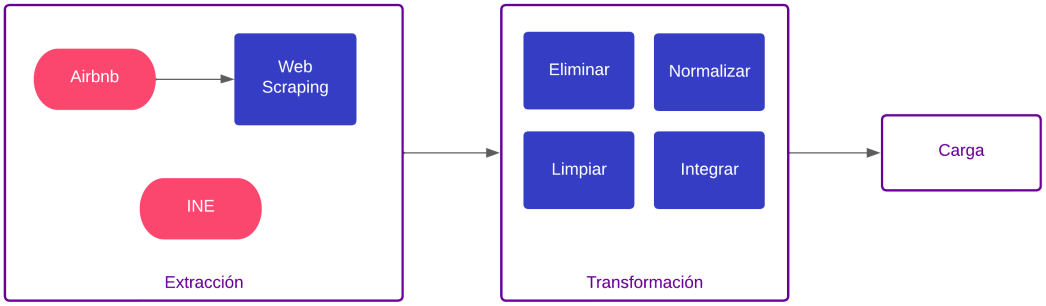

Figura 6.2: Proceso ETL del proyecto

CSV, por lo que se ha procedido a la descarga de aquéllos relevantes para la realización de este proyecto.

Si bien es cierto que el acceso a estos datos es sencillo, la distribución de los mismos en los archivos CSV no es el más adecuado para trabajar con ellos. Como se puede apreciar en la Tabla 6.1, la información es bastante redundante y poco comprensible. Más adelante se verá cómo se ha adaptado a nuestras necesidades e integrado en las diferentes tablas.

Comunidades	Gasto y duración	Tipo de dato	Periodo	Total
01 Andalucía	Gasto total	Dato base	2022	11989,59
01 Andalucía	Gasto total	Dato base	2021	4768,27
01 Andalucía	Gasto total	Dato base	...	...
01 Andalucía	Gasto total	Dato base	2016	11318,92
01 Andalucía	Gasto total	Tasa de variación anual	2022	151,45
01 Andalucía	Gasto total	Tasa de variación anual	2021	65,85
01 Andalucía	Gasto total	Tasa de variación anual	...	...
01 Andalucía	Gasto total	Tasa de variación anual	2016	8,14
01 Andalucía	Gasto medio por persona	Dato base	2022	1198
01 Andalucía	Gasto medio por persona	Dato base	2021	1122
...	...	...	...	...

Tabla 6.1: Datos INE sin procesar

Extracción de datos de Airbnb

Por otro lado, respecto a la aplicación de Airbnb, la información debe extraerse de cada página de alojamiento única, siendo necesaria la utilización, en este caso, de técnicas de Web Scraping. Para el desarrollo de dichas técnicas ha sido necesaria la implementación de código propio. Los recursos disponibles en la Web para este tipo de tareas, como pueden ser extensiones de navegadores o aplicaciones en línea, no satisfacían las necesidades específicas del

trabajo, pues no ofrecían la flexibilidad y personalización necesarias para extraer los datos de Airbnb de manera eficiente y precisa. Entre las herramientas que se han intentado utilizar se encuentran Octoparse, la cuál no ofrecía las funcionalidades necesarias para obtener los datos deseados y Algolia Crawler, que en su versión gratuita presentaba demasiadas limitaciones. En consecuencia, se optó por desarrollar código propio utilizando lenguajes de programación como Python y aprovechando bibliotecas especializadas, como BeautifulSoup o Selenium, que proporcionan funcionalidades robustas para analizar y extraer información de páginas web. Además, éste se complementó con la ayuda de una librería de GitHub [24]. Dicha librería permitía la interacción con Airbnb pero, debido al tiempo transcurrido desde su creación, fue necesaria la implementación de cambios sobre el código original. Las modificaciones realizadas posibilitaron interactuar con las páginas de Airbnb y extraer la información relevante con mayor flexibilidad y control.

En los siguientes subapartados se describen las diferentes fases que han sido necesarias en este proceso:

Obtención de enlaces

La primera tarea a llevar a cabo era, como cabe esperar, la obtención de los diferentes enlaces de los alojamientos y extracción de sus características. Para la realización de este punto se han creado una serie de enlaces a las diferentes localizaciones de nuestro interés. En el caso de España, se ha hecho una búsqueda de apartamentos por provincia. Se usó esta estrategia debido a las limitaciones que impone la plataforma a la hora de acceder a la misma, ya que si detecta demasiadas peticiones en un plazo corto de tiempo bloquea su acceso durante días.

Para el correcto desempeño de de esta tarea, se ha partido de un Data Frame con todas las provincias del país y se han recorrido cada una de ellas e insertado en el enlace de referencia de la plataforma (https://www.airbnb.es/s/provincia/homes) como se muestra a continuación.

```
#Generamos los enlaces para cada una de las provincias
def generar_enlaces_airbnb(nombre_archivo):
    # Leer el archivo XLS
    df = pd.read_excel(nombre_archivo)
    # Crear una lista vacía para almacenar los enlaces de Airbnb
    enlaces_airbnb = []

    # Recorrer la columna de los nombres de los municipios y
    generar los enlaces correspondientes
    for provincia in df['NOMBRE']:
        enlace =
    'https://www.airbnb.com/s/{}/homes'.format(provincia)
        enlaces_airbnb.append(enlace)
```

```
12
13      return enlaces_airbnb
```

Cada uno de estos enlaces devuelve la búsqueda de la provincia correspondiente en la plataforma. En cada búsqueda aparecerán por pantalla varias pestañas con diversos alojamientos cada una. Se trata de obtener los enlaces a cada uno de estos alojamientos. Para poder acceder a las diferentes pestañas de una búsqueda debe agregarse el parámetro *items_offset* al enlace. Luego, en cada una de ellas podremos extraer las URLs de los diferentes apartamentos procesando los códigos HTML como sigue:

```
1  #Esta función además de para buscar más adelate otros elementos,
       ahora nos vale para dentro de las listas de la siguiente
       función, poder solamente el enlace del contenido obtenido
2  def extract_element(listing_html, params):
3       # Busca el elemento en función de la etiqueta y la clase
       especificadas en el diccionario params
4      if 'class' in params:
5          elements_found = listing_html.find_all(params['tag'],
       params['class'])
6      else:
7          elements_found = listing_html.find_all(params['tag'])
8
9      # Seleccionamos el elemento específico que se devolverá en
       función del parámetro order, si no devolverá el primer elemento
10     tag_order = params.get('order', 0)
11     element = elements_found[tag_order]
12
13     # Extraemos el texto
14     if 'get' in params:
15         output = element.get(params['get'])
16     else:
17         output = element.get_text()
18
19     return output
20
21 #Extraemos el enlace de los alojamientos usando las bibliotecas
       "requests", "BeautifulSoup" y la función extract_element() y
22 #devolvemos una lista con los diferentes enlaces.
23 def get_htmls(search_page):
24     answer = requests.get(search_page, timeout=30) # envío
       solicitud HTTP GET a la URL
25     content = answer.content # obtiene el contenido de la respuesta
26     soup = BeautifulSoup(content, 'html.parser') #analiza el
       contenido HTML y representa el árbol de elementos HTML.
27     listings = soup.find_all('div', 'cy5jw6o') #busca todos los
       elementos HTML que tienen la etiqueta "div" y la clase
```

```
    "cy5jw6o", esta clase contiene los enlaces de los apartamentos
    href_list = []
    for listing in listings:
            features = extract_element(listing, {'url': {'tag':
    'a', 'get': 'href'}}) # se llama "extract_page_features" con el
    elemento y un diccionario que especifica la regla para extraer
    la URL de ese elemento HTML
            href_list.append(features) # agrega cada conjunto de
    características extraídas a una lista de características
    "features_list"

    return href_list
```

En este fragmento se puede ver cómo es necesario comprender el código HTML para extraer del mismo las etiquetas relevantes y obtener la información deseada en cada momento. En este caso, en concreto, se extrae la URL que vendrá recogida en el tag 'a', tras un indicador 'href', el cual indicará que le sigue una URL.

Obtención de características

Una vez se han obtenido todas las URLs de los alojamientos de interés se procede a la extracción de sus características. Para llevar a cabo esta tarea se ha hecho uso de la librería Beautiful Soup y se han establecido las reglas de búsqueda en función de las etiquetas HTML a través del siguiente diccionario:

```
RULES_DETAIL_PAGE = {
    'name': {'tag': 'span', 'class': '_1n81at5'},

    'habitaciones':{'tag': 'li', 'class': 'l7n4lsf', 'order': -1},

    'viajeros': {'tag': 'li', 'class': 'l7n4lsf'},

    'amenities': {'tag': 'div', 'class': '_1qsawv5', 'order': -1},

    'price_per_night': {'tag': 'span', 'class': '_tyxjp1'},

    'cancelacion': {'tag': 'div', 'class': 'tutyap8'},

    'media_ratings': {'tag': 'span', 'class': '_17p6nbba'},

    'n_evals': {'tag': 'li', 'class': 'tpa8qb9', 'order':-1},
    'host_feats': {'tag': 'ul', 'class': 'fhhmddr', 'order': -1},

    'house_rules': {'tag': 'div', 'class': 'is50c2g', 'order': -1},
}
```

En este diccionario cada clave representa el nombre del atributo que se desea extraer y su valor es otro diccionario con información sobre cómo encontrar y obtener esos elementos. Por ejemplo, para el atributo 'name' se especifica la clase '_1n81at5' del elemento para extraer el nombre. Con la definición de estas reglas se han podido encontrar y extraer las características de interés de cada alojamiento.

Se debe tener en cuenta que estas reglas cambian a menudo en el tiempo; a lo largo mismo de la realización del proyecto la página web fue actualizada y estas reglas se han visto modificadas para poder proseguir con la extracción.

A mayores, cabe mencionar que la extracción de latitud y longitud ha sido más compleja por el hecho de que la propia aplicación no muestra, de entrada, la ubicación exacta, y la ubicación aproximada se encuentra dentro de un mapa en la propia página. El fragmento de código creado para la extracción de coordenadas es:

```
1
2  scripts = soup.find_all("script") #soup se corresponde con el
       elemento de árbol "BeautifulSoup" que permite acceder y
       manipular fácilmente los elementos y datos de la página web
       utilizando métodos y atributos proporcionados por BeautifulSoup.
3
4  for script in scripts:
5      if '"lat"' in str(script):
6          script_content = script.text
7
8
9  # Extraer los valores de las variables 'lat' y 'lng' del contenido
       del script
10 if script_content:
11     lat_index = script_content.index('"lat":') + 6
12     lng_index = script_content.index('"lng":') + 6
13     lat = script_content[lat_index:lat_index+5]
14     lng = script_content[lng_index:lng_index+5]
15 else:
16     lat = None
17     lng = None
```

Una vez obtenidos los datos de las páginas individuales, se han almacenado en una estructura de datos adecuada, un archivo CSV, para su posterior análisis. En la Tabla 6.2 se pueden ver las características extraídas para dos de los alojamientos.

Características	Alojamiento 1	Alojamiento 2
name	Apartamento de lujo 1ª linea de Playa Finisterre	Loft Finisterre Centro
price_per_night	59 €	78 €
habitaciones	2 dormitorios · **__** · 2 camas · **__** · 2 baños**__**Se registró en junio de 2012 ·	1 dormitorio · **__** · 1 cama · **__** · 1 baño**__**Se registró en febrero de 2019 ·
viajeros	3 viajeros ·	2 viajeros ·
amenities	”amenity_lavander00eda”, ”amenity_Internet y oficina”, ”amenity_Aparcamiento”, …	”amenity_lavander00eda”, ”amenity_Calefacci00f3n y refrigeraci00f3n”, …
cancelacion	Esta reserva no es reembolsable.	Esta reserva no es reembolsable.
media_ratings	4,82 ·	4,72 ·
n_evals	202 evaluaciones**__**��Superanfitrión	120 evaluaciones**__**��Superanfitrión
host_feats	Llegada a partir de las 15:00**__**Salida antes de las 11:00**__**Máximo 6 huéspedes**__**Lago, río u otra masa de agua cerca	Llegada a partir de las 16:00**__**Salida antes de las 11:00**__**Máximo 2 huéspedes
lat	42.91	42.90
lng	-9.26	-9.26
mes_extraccion	5	5
ano_extraccion	2023	2023

Tabla 6.2: Datos extraídos de Airbnb sin procesar

Este proceso de extracción se ha demorado, como se ha indicado en la planificación, más de lo previsto, debido a que la información que debía extraerse se encontraba alojada en cada una de las páginas de alojamiento únicas, por lo que se ha tenido que navegar por cada una de las ellas y extraer las características de forma individual. A ello hay que añadir el hecho de que se trata de una tarea con un alto coste de recursos, tanto temporales como computacionales, y que, además, se ha visto afectada por problemas adicionales, como han sido el acceso denegado por parte de la plataforma a consecuencia de la detección de robots.

6.2.2 Limpieza de los datos

La limpieza de datos es una etapa crucial para obtener un buen modelo de datos y poder realizar, de manera óptima, un análisis de los mismos, garantizando su fiabilidad y coherencia.

Limpieza de datos de Airbnb

Mediante el uso de técnicas de *Web scraping* los datos obtenidos pueden contener ruido, errores, formatos extraños e incluso duplicados (porque un apartamento salga en diferentes búsquedas, por ejemplo). Esta variedad de problemas compromete la calidad de los resultados e incrementa la dificultad de las tareas posteriores.

Para solucionar dichos problemas se han utilizado diversas estrategias. Como se puede

observar en la Tabla 6.2 hay columnas que simplemente se deben convertir a un formato adecuado ('price_per_night' o 'viajeros') para poder trabajar con ellas, pero también existen otras en las que encontramos más datos de los previstos y que, además, hay que adaptar ('amenities' o 'habitaciones').

En la primera situación, simplemente basta con almacenar el valor numérico o sustituir el valor por otro más sencillo. A continuación se ve cómo tratar cada uno de los casos:

- **price_per_night, media_ratings y viajeros:** En estos tres casos lo que interesa de estas características es su valor numérico por tanto solo se necesita extraer el mismo y eliminar el resto. Este proceso se ha llevado a cabo con la ayuda del lenguaje Python y expresiones regulares, como se ve en el ejemplo que sigue:

```
df['media_ratings'] = df['media_ratings'].str.rstrip('· ')

df['price_per_night'] =
df['price_per_night'].str.extract(r'(\d+\.?\d*)€?')

```

- **cancelacion:** En el caso de la 'cancelación' tenemos tres posibles opciones, que se reembolse el dinero completamente, que no se reembolse o que se reembolse parcialmente. Una vez analizadas las diferentes opciones que se nos presentan en esta columna se procede de manera similar a la anterior, y se normalizan los valores con Python y la ayuda de las expresiones regulares:

```
df['Cancelacion'] = np.select(
  [
      df['cancelacion'].str.contains('no es reembolsable',
case=False),
      df['cancelacion'].str.contains('cancelación gratuita',
case=False),
      df['cancelacion'].str.contains('reembolso parcial',
case=False)
  ],
  [
      'no_reembolsable',
      'reembolsable',
      'parcialmente_reembolsable'
  ],
  default= 'NaN'
)

```

En la segunda situación, en la cual hay muchos datos en una misma columna, se procede de una de las siguientes dos maneras:

- Eliminando los datos que no interesen de la columna y preservando los que sí.
- Dividiendo la columna en otras nuevas, con sus diferentes valores (en caso de que sean de interés).

A continuación, se ven las diferentes situaciones que se han presentado:

- **habitaciones, n_evals y host_feats:** En estos campos no aparece una sola característica sino que se presentan varias; por lo que se procede a organizarlas en columnas independientes. Para ello se extraen inicialmente los valores de interés, nuevamente con expresiones regulares, y se colocan en sus columnas correspondientes. A continuación, se presenta el ejemplo del campo 'habitaciones'; en este caso los dormitorios, camas y baños vendrán representados por su valor numérico y el registro en mes y año (se procede de manera similar para las otras tres columnas):

```
df['Dormitorios'] =
df['habitaciones'].str.extract(r'(\d+)\s*dormitorios?')
.astype(float)
df['Camas'] =
df['habitaciones'].str.extract(r'(\d+)\s*cama?').astype(float)
df['Baños'] =
df['habitaciones'].str.extract(r'(\d+)\s*baño?').astype(float)
df['Fecha_registro'] = df['habitaciones'].str.extract(r'Se
registró en(.*?)\*\*__\*\*')
# Extraer el mes y el año de la columna 'fecha_registro' y
guardarlos en dos nuevas columnas
df['mes_registro'] =
df['Fecha_registro'].str.extract(r'(enero|febrero|marzo|abril|

mayo|junio|julio|agosto|septiembre|octubre|noviembre|diciembre)')
df['ano_registro'] =
df['Fecha_registro'].str.extract(r'(\d{4})')

df.drop(['habitaciones', 'Fecha_registro'], axis=1,
inplace=True)

```

- **house_rules:** En este caso, además de contar con datos a almacenar como puede ser el horario de llegada y salida, hay también información redundante o que no interesa almacenar. El caso de *'máximo 6 huéspedes'* es redundante, ya que si la capacidad máxima del alojamiento (dada por 'viajeros') ya se ha visto que son 6 huéspedes, se entiende que una norma será que no se aloje una cantidad mayor de usuarios. Por otro lado, también se elimina ruido como es *'Lago, río u otra masa de agua cerca'*. El procedimiento para

limpiar esta columna es análogo al anterior; se crean nuevas columnas para los horarios y, a continuación, se borra 'house_rules'.

```
df['Hora_entrada'] = df['house_rules'].str.extract('Llegada a
partir de las (\d+:\d+)', expand=True)
df['Hora_salida'] = df['house_rules'].str.extract('Salida antes
de las (\d+:\d+)', expand=True)

df.drop(['house\_rules'], axis=1, inplace=True)

```

- **amenities:** Para esta columna simplemente se ha eliminado el prefijo 'amenity_' y mantenido el resto.

```
df['amenities'] = df['amenities'].apply(lambda x: ',
'.join([i.replace('amenity_','') for i in eval(x).keys()]))

```

Además, se vio interesante obtener los municipios de cada uno de los apartamentos a partir de su latitud y su longitud, para lo cual se necesitó la librería Geopandas, un archivo GeoJSON [20] con los municipios, y convertir nuestro DataFrame en un GeoDataFrame:

```
import geopandas as gpd

#Cargar el archivo GeoJSON con los municipios
municipios =
    gpd.read_file('georef-spain-municipioyprovincia.geojson')

# Convertir el DataFrame en un objeto GeoDataFrame
gdf = gpd.GeoDataFrame(df, geometry=gpd.points_from_xy(df.lng,
    df.lat))

# Realizar la intersección entre los puntos y los polígonos
    (municipios)
df = gpd.tools.sjoin(gdf, municipios, op='within')

#Nos quedamos simplemente con las columnas que nos interesan
df.drop(['geo_point_2d', 'year',  'mun_area_code','mun_type',
    'mun_name_local', 'geometry'], axis=1, inplace=True)
```

Ya para finalizar, y antes de dividir esta gran tabla en las diferentes dimensiones y hechos a considerar, se realizó una comprobación de que no hubiese registros de alojamientos duplicados que pudiesen interferir en el posterior estudio de los mismos y se les asignó un identificador único a cada uno:

```
# Eliminar los duplicados y mantener solo una instancia de cada
    fila duplicada
df = df.drop_duplicates()
# Verificar si hay filas duplicadas
duplicados = df.duplicated()
# Si no hay filas duplicadas, agregar columna de ID
if not duplicados.any():
  df['id_alojamiento'] = df.reset_index().index
```

La tabla con todas las características de Airbnb, para los dos alojamientos de muestra, quedó finalmente como se puede observar en la Tabla 6.3.

Características	Alojamiento 1	Alojamiento 2
id_alojamiento	0	1
name	Apartamento de lujo 1ª linea de Playa Finisterre	Loft Finisterre Centro
price_per_night	59 €	78 €
Dormitorios	2	1
Camas	2	1
Baños	2	1
mes_registro	junio	febrero
ano_registro	2012	2019
viajeros	3	2
amenities	Lavandería, Internet y oficina, Aparcamiento, ...	Lavandería, Calefacción y refrigeración, ...
cancelacion	no_reembolsable	no_reembolsable
media_ratings	4,82	4,72
n_evals	202	120
superanfitrion	si	si
idiomas	English, Italiano, Español	empty
ratio_respuesta	100%	100%
tiempo_respuesta	en menos de una hora	en menos de una hora
hora_llegada	15:00	16:00
lat	42.91	42.90
lng	-9.26	-9.26
id_mun	2709	2709
mun_name	Fisterra	Fisterra
prov_code	15	15
prov_name	A Coruña	A Coruña
id_ccaa	12	12
ccaa_name	Galicia	Galicia
mes_extraccion	5	5
ano_extraccion	2023	2023

Tabla 6.3: Datos extraídos de Airbnb procesados

Una vez se tuvieron todas las características extraídas procesadas resultó de interés añadir algunas a mayores, a partir de las que ya se tenían, para hacer su análisis más sencillo. En el caso de las 'amenities' y de los 'idiomas' se decidió incluir columnas con el número de valores de cada una para poder trabajar con ellas de manera más óptima.

```
# Actualizar el valor de 'num_amenities' en el DataFrame
df['num_amenities'] = df['amenities'].apply(lambda x: 0 if
    pd.isnull(x) or x.strip() == '' else len(x.split(',')))

# Actualizar el valor de 'num_idiomas' en el DataFrame
df['num_idiomas'] = df['idiomas'].apply(lambda x: len(x.split(','))
    if isinstance(x, str) else 0)
```

Además, para las 'amenities' se tomó la decisión de crear cuatro columnas con las cuatro más relevantes (aparcamiento, internet, calefacción y vistas panorámicas), en las cuales se indica si el alojamiento cuenta o no con ese servicio. A continuación, vemos el ejemplo de 'internet':

```
    # Definimos una función que toma el valor de la columna
    'amenities' y devuelve 'Si' si contiene 'Internet', y 'No' en
    caso contrario.
def internet_check(amenities):
    if isinstance(amenities, str) and 'Internet' in amenities:
        return 'Si'
    else:
        return 'No'

# Aplicamos la función a la columna 'amenities' y guardamos los
    resultados en una nueva columna llamada 'Internet'
df['Internet'] = df['amenities'].apply(Internet)
```

Hecho esto, el siguiente paso es obtener las diferentes tablas que conformarán nuestro almacén.

- **Dimensión anfitrion y cancelacion:** Para obtener estas dimensiones, sabiendo las opciones de valor que podían tener, se han creado las tablas con sus correspondientes valores y, seguidamente, se les ha asignado un identificador. A continuación, se muestra el código perteneciente a la dimensión 'cancelación' (con la de 'anfitrión' se procede de la misma manera).

```
    #Creamos un df para los tipos de anfitriones
    # Crear el DataFrame
    data = {'id_cancelacion': [0, 1, 2, 3],
                'tipo_cancelacion': ['no_reembolsable',
    'reembolsable', 'reembolsable_parcialmente', 'NaN']}

```

```
    df_cancelacion = pd.DataFrame(data)

    df_cancelacion.to_csv('cancelacion.csv')

    # Definir una función para asignar los valores a la columna
    'id_cancelacion'
    def asignar_id_cancelacion(row):
        if row['Cancelacion'] == 'no_reembolsable':
            return 0
        elif row['Cancelacion'] == 'reembolsable':
            return 1
        elif row['Cancelacion'] == 'parcialmente_reembolsable':
            return 2
        else:
            return 3

    # Aplicar la función a cada fila del DataFrame para obtener los
    valores de 'id_anfitrion'
    df['id_cancelacion'] = df.apply(asignar_id_cancelacion, axis=1)
    df.drop(['Cancelacion'], axis=1, inplace=True)

```

- **Dimensión fecha:** En el caso de la creación de fechas se ha optado por ir a la fecha más antigua, Febrero de 2010, y almacenar cada mes de cada año con un identificador.

```
    # Crear una lista de todos los meses desde febrero de 2010
    hasta junio de 2023
    start_date = pd.to_datetime('2010-02-01')
    end_date = pd.to_datetime('2023-06-01')
    date_range = pd.date_range(start=start_date, end=end_date,
    freq='MS')

    # Crear un DataFrame vacío
    df_fecha = pd.DataFrame(columns=['mes', 'año', 'identificador'])

    # Enumerar el identificador iniciando desde 0
    identifier_counter = 0

    # Llenar el DataFrame con los meses, años e identificadores
    for date in date_range:
        month = date.month
        year = date.year
        identifier = identifier_counter
        identifier_counter += 1
        df_fecha = df_fecha.append({'mes': month, 'año': year,
    'identificador': identifier}, ignore_index=True)
```

```
    df_fecha.to_csv('fecha.csv')
```

- **Dimensión alojamientos:** Para obtener esta dimensión simplemente se han seleccionado las características correspondientes de la tabla inicial y almacenado en un archivo CSV para su posterior carga:

```
    df_alojamientos = df[['id_alojamiento', 'name', 'viajeros',
    'media_ratings', 'n_evals', 'amenities', 'lat', 'lng',
    'Dormitorios', 'Camas',
    'Baños', 'Idiomas', 'Ratio_respuesta',
    'Tiempo_respuesta', 'hora_llegada', 'Hora_salida', 'registro',
    'id_mun', 'mes_registro', 'año_registro', 'num_idiomas',
    'num_amenities', 'calefaccion_refrigeracion', 'internet',
    'vistas_panoramicas', 'aparcamiento']]

    df_alojamientos.to_csv('alojamientos.csv')
```

- **Dimensión municipios y ccaa:** En el caso de los municipios y comunidades autónomas se ha procedido de forma similar, pero eliminando los duplicados, ya que en este caso, al hacer su extracción de la tabla, existen muchos alojamientos en un mismo municipio, y por tanto, éste aparecerá varias veces:

```
    df_mun = df[['id_mun', 'mun_name', 'prov_code', 'prov_name',
    'id_ccaa']]
    df_mun = df_mun.drop_duplicates()
    df_mun.to_csv('mun.csv')

    df_ccaa = df[['id_ccaa', 'ccaa_name']]
    df_ccaa = df_ccaa.drop_duplicates()
    df_ccaa.to_csv('ccaa.csv')
```

- **Hecho alquiler:** Finalmente, se procede igual para la tabla de hechos, seleccionando las columnas pertinentes del DataFrame:

```
    df_alquiler = df[['id_alojamiento', 'price_per_night',
    'id_anfitrion', 'id_cancelacion', 'mes_extraccion',
    'ano_extraccion']]
    df_alquiler.to_csv('mun.csv')
```

Limpieza de datos del INE

En el caso de los datos ofrecidos por el INE, como ya se ha comentado anteriormente, además de venir en tablas diferentes, su formato no era el más adecuado. En esta situación, se optó por integrarlos ordenados por años y comunidades, en una misma tabla, con la ayuda del lenguaje Python (ver ejemplo en Tabla 6.4):

```
for index, row in gasto.iterrows():
    if row['Gastos y duración media de los viajes'] == 'Gasto
    total' and row['Tipo de dato'] == 'Dato base':
        gasto_total.append((row['Total']))
    if row['Gastos y duración media de los viajes'] == 'Gasto medio
    por persona' and row['Tipo de dato'] == 'Dato base':
        gasto_medio_persona.append(row['Total'])
    if ...
```

Como se puede observar en el fragmento de código anterior, se ha ido iterando sobre las filas del DataFrame realizando comprobaciones para determinar qué columnas y valores se debían extraer y almacenar. Una vez hecho esto, se ha creado el DataFrame resultante y asignado el identificador correspondiente a cada Comunidad Autónoma.

id_ccaa	Periodo	gasto_total	gasto_medio_persona	...
1	2022	11989,59	1198	...
1	2021	4768,27	1122	...
...	...	...	...	...

Tabla 6.4: Datos INE procesados

A mayores, en relación a los gastos totales recogidos por meses y años, cabe indicar que estos datos ya se suministraron de manera ordenada (Tabla 6.5), por lo que simplemente se procedió a eliminar algunas columnas sobrantes, como las diferentes variables y valores numerados, y a separar la columna correspondiente al periodo, en su respectivo mes y año.

```
    # Extraer el mes y el año en columnas separadas
    df['mes'] = df['PERIODO'].str.extract(r'M(\d{2})')
    df['año'] = df['PERIODO'].str.extract(r'(\d{4})')
```

Variable1	Valor1	...	PERIODO	VALOR
Tipo de dato	Dato Base	...	2023M04	8479,99
Tipo de dato	Dato Base	...	2023M03	6657,00
...	...	...	...	...

Tabla 6.5: Gastos por mes y año sin procesar

Esto dio como resultado la Tabla 6.6

mes	año	VALOR
4	2023	8479,99
3	2023	6657,00
...	...	...

Tabla 6.6: Gastos por mes y año procesados

Por último, y antes de proceder con la carga de datos, se asignaron los identificadores correspondientes de la dimensión 'fecha' a cada tabla, realizando una serie de comparaciones. Por ejemplo, para la dimensión de 'alojamientos' se comprobó que coincidiesen los campos 'mes_registro' y 'año_registro' con el 'mes' y el 'año' de la dimensión 'fecha'.

```
UPDATE alojamientos
INNER JOIN fecha ON alojamientos.año_registro = fecha.año AND
alojamientos.mes_registro = fecha.mes
SET alojamientos.id_fecha_registro = fecha.id_fecha;
```

El mismo proceso se llevó a cabo con las tablas de 'alquiler' ('año_extraccion' y 'mes_extraccion'), 'gasto_total' ('mes' y 'año') e 'ine' ('periodo'; el mes a asignar es siempre 12, ya que son valores que hacen referencia al año entero). Posteriormente, se eliminaron las columnas relativas al mes y al año dejando solo los identificadores.

6.2.3 Carga de los datos

La última fase ETL, supone la carga de datos en el almacén para tenerlos a nuestra disposición y poder explotarlos.

Creación de la base de datos

Antes de la carga, y habiendo ya definido previamente la estructura de lo que será el Data Warehouse, un último paso a realizar es la creación de la base de datos y de las tablas correspondientes a las diferentes dimensiones y hechos. Para ello se ha utilizado el sistema gestor de base de datos relacional MySQL.

La creación de la base de datos se realizó ejecutando el siguiente comando SQL:

```
CREATE DATABASE airbnb;
```

Respecto a la creación de las diferentes tablas se procedió de igual forma para todas ellas. En el siguiente ejemplo se puede ver el código ejecutado para la creación de la tabla correspondiente a la dimensión 'alojamientos'.

```
CREATE TABLE `alojamientos` (
  `id_alojamiento` int NOT NULL,
  `name` varchar(250) DEFAULT NULL,
```

```
  `viajeros` double DEFAULT NULL,
  `media_ratings` double DEFAULT NULL,
  `n_evals` int DEFAULT NULL,
  `amenities` varchar(300) DEFAULT NULL,
  `lat` double DEFAULT NULL,
  `lng` double DEFAULT NULL,
  `Dormitorios` double DEFAULT NULL,
  `Camas` double DEFAULT NULL,
  `Baños` double DEFAULT NULL,
  `Idiomas` varchar(200) DEFAULT NULL,
  `Ratio de respuesta` varchar(4) DEFAULT NULL,
  `Tiempo de respuesta` varchar(20) DEFAULT NULL,
  `Hora_entrada` time DEFAULT NULL,
  `Hora_salida` time DEFAULT NULL,
  `id_mun` int DEFAULT NULL,
  `año_registro` int DEFAULT NULL,
  `mes_registro` int DEFAULT NULL,
  `num_amenities` int DEFAULT NULL,
  `num_idiomas` int DEFAULT NULL,
  `internet` varchar(3) DEFAULT NULL,
  `vistas_panoramicas` varchar(3) DEFAULT NULL,
  `calefaccion_refrigeracion` varchar(3) DEFAULT NULL,
  `aparcamiento` varchar(3) DEFAULT NULL,
  PRIMARY KEY (`id_alojamiento`))
```

El script utilizado para la creación de cada una de las tablas se puede encontrar en el Apéndice A.

Una vez creadas las tablas, se procedió a la creación de relaciones entre ellas mediante la definición de claves foráneas.

```
-- Crear la restricción de clave foránea en la columna 'columna' de
    la tabla 'nombre_tabla'
ALTER TABLE nombre_tabla
ADD CONSTRAINT fk_name
FOREIGN KEY (columna)
REFERENCES tabla_relacion(columna);
```

Procediendo de esta manera para cada una de las relaciones se obtuvo finalmente el esquema deseado. El script utilizado para la creación de cada una de las relaciones se puede encontrar en el Apéndice B.

Carga de datos con Talend

Creadas las tablas y establecidas las relaciones de la base de datos, se implementó un proceso de carga de datos haciendo uso de Talend para automatizar el proceso y que éste

fuese escalable e incremental. Este enfoque permite tener una solución robusta para manejar grandes volúmenes de información y facilitar futuras actualizaciones de los datos. Dicha plataforma ofrece, también, soporte para la limpieza y transformación de los datos; en este caso se optó por usar Python, debido a la familiarización con el lenguaje y su amplia flexibilidad en la ejecución de este tipo de tareas, pero el uso de Talend también se podría considerar una opción.

Como paso inicial del proceso de carga se utilizó el componente de Talend 'tFileInputDelimited' para obtener la lista de archivos CSV correspondientes a cada una de las tablas y leer su contenido.

Para poder comenzar con la carga de datos hay que tener claros los puntos a cubrir y tratar. Dicha carga se basará en crear conexiones entre los archivos CSV, con la información extraída y procesada, y la base de datos. Este proceso debe dar soporte a posibles actualizaciones incrementales, las cuales Talend nos resuelve con la herramienta 'tMap'. Estas actualizaciones se corresponderán con cambios, por ejemplo en las características de los alojamientos o alquileres, y la adición de información a futuro.

A continuación se describe la carga de cada una de las dimensiones y tablas de hechos.

Dimensión Cancelación

Se trata de una dimensión que no necesitará de actualizaciones ni inserciones futuras; para llevar a cabo su carga de datos basta con conectar el origen de datos, el archivo CSV, con el destino, la tabla correspondiente de la base de datos, como se puede ver en la Figura 6.3.

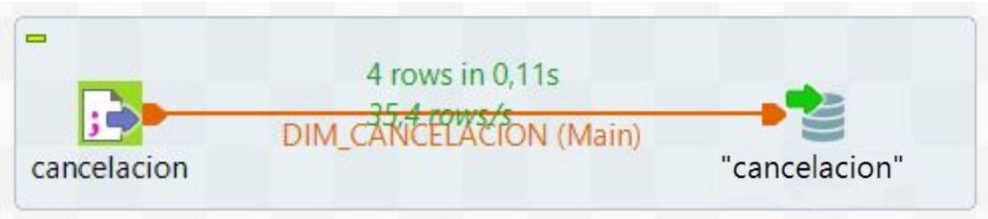

Figura 6.3: Carga Dimensión Cancelación

Dimensión Anfitrión

De la misma manera que en el caso anterior se procede también en este caso; visualizamos este procedimiento en la Figura 6.4.

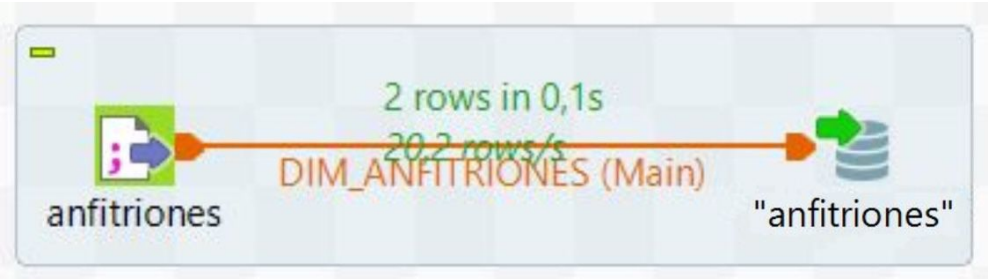

Figura 6.4: Carga Dimensión Anfitrión

Dimensiones CCAA y Municipios

Siguiendo la misma línea previa, con estas dimensiones se ha procedido de igual manera, ya que el número de comunidades autónomas y municipios no es información que vaya a variar o precisar de actualización. Se puede observar el proceso de carga en las figuras Figura 6.5 y Figura 6.6.

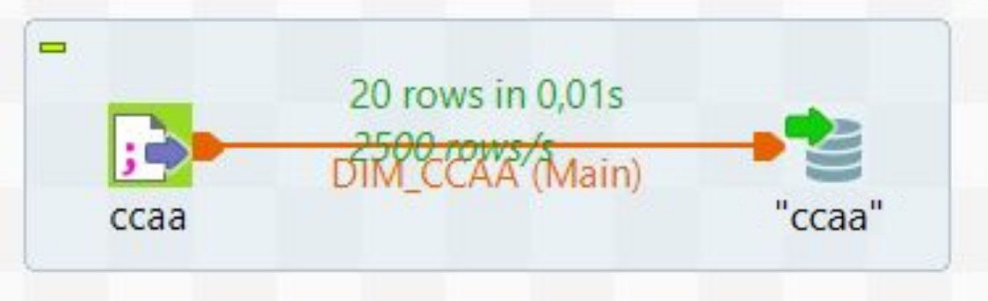

Figura 6.5: Carga Dimensión CCAA

Figura 6.6: Carga Dimensión Municipios

Dimensión Fecha

En el caso de los datos relativos a las diferentes fechas ocurre algo diferente; según se avanza en el tiempo se tendrán que añadir cada uno de los meses de cada uno de los años. Esto se debe, entre otros, a que se obtendrá nueva información referida al total de gastos, habrá nuevos registros de alojamientos en la plataforma, se podrán hacer nuevos procesos de extracción en otras fechas, etc.

Para tratar esta incorporación sin tener que volver a cargar los datos ya incorporados con anterioridad se ha hecho uso del componente 'tMap'. Dicho componente permite llevar a cabo la carga con una serie de requisitos específicos que le indiquemos, lo cual hace que esta herramienta sea fundamental para asegurar la eficiencia del proceso. En la Figura 6.7 podemos ver el mapeo de datos realizado en esta dimensión, y cómo se introdujo una expresión para que solo se inserten datos en el caso de que su identificador no se encuentre en la base de datos. Para poder realizar dicho mapeo éste debe recibir como entrada tanto el CSV de los datos a cargar como el propio acceso a los datos ya cargados en la base de datos.

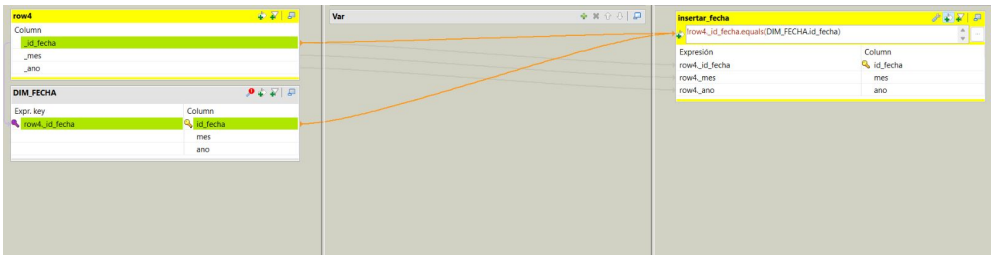

Figura 6.7: tMap Dimensión Fecha

En la Figura 6.8 se puede ver la carga inicial, cuando la base de datos aún se encuentra vacía. Por otro lado, en la Figura 6.9 se muestra cómo si se intentan insertar datos que ya están, estos no se insertan de nuevo, solo se hará en el caso de que no existan.

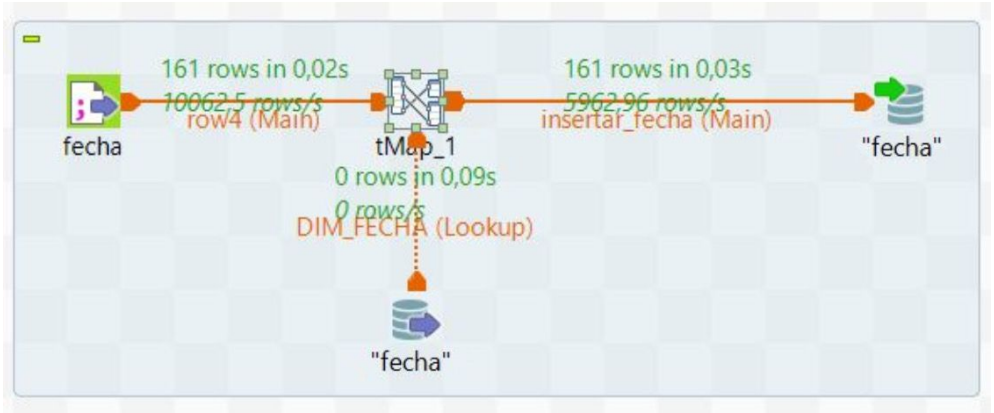

Figura 6.8: Carga Dimensión Fecha

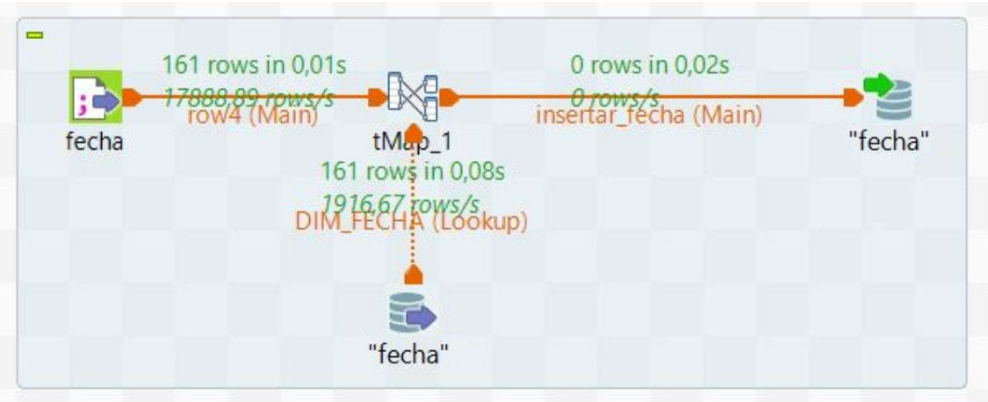

Figura 6.9: Insertar datos duplicados en la Dimensión Fecha

Dimensión Alojamientos

Esta dimensión tiene que permitir actualizaciones en sus datos. Esto es, si un alojamiento, por ejemplo, se llama 'Casa Lucía' y su dueña decide cambiar su nombre y poner 'Casa en el mar', este dato debe actualizarse en la base de datos. Además, se debe contar con la opción, también, de insertar nuevos alojamientos.

Nuevamente se ha echado mano del componente 'tMap' para definir las diferentes condiciones y que, según el dato, se proceda a una inserción, una actualización o bien no se haga nada. En este caso las entradas son dos, el CSV y la correspondiente dimensión de la base de datos, teniéndose, además, dos posibles salidas, la de inserción y la de actualización. El esquema y la carga inicial de esta dimensión se puede ver en la Figura 6.10.

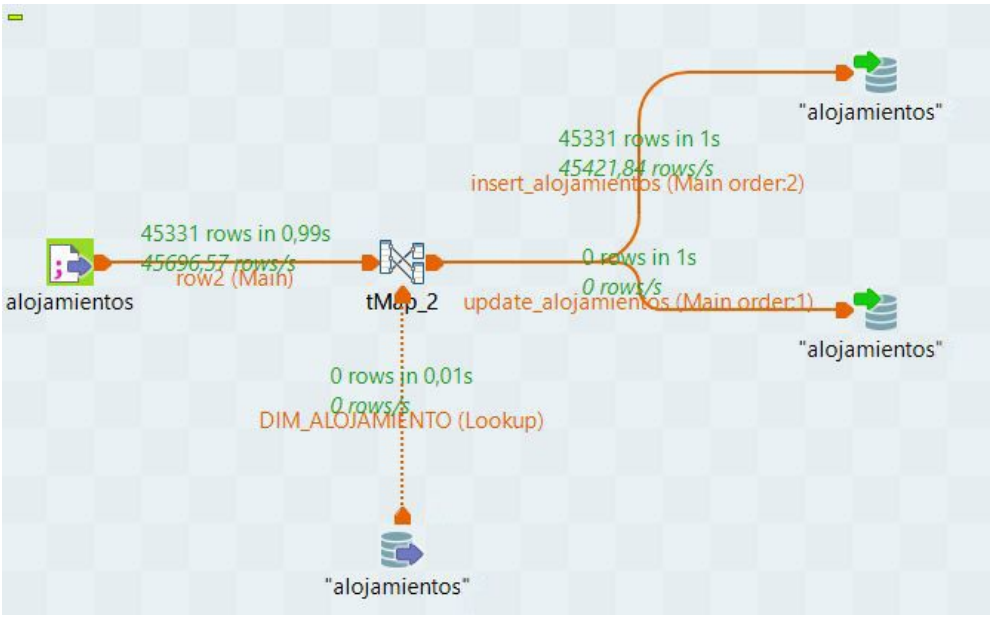

Figura 6.10: Carga Dimensión Alojamientos

Por otro lado, el mapeo de los datos se muestra en la Figura 6.11 (la expresión completa de la actualización se encuentra en la Figura 6.12). Como se observa, la inserción se realiza como sucede con la fecha, solo en el caso de que el identificador aún no esté en la base de datos, mientras que la actualización se efectúa cuando este identificador ya existe y, además, se detecta un cambio en alguna de las columnas indicadas en la expresión. Asimismo, para que dicha actualización modifique los datos de la base de datos y no inserte una nueva fila, se debe marcar el valor de la acción 'Actualizar', como se observa en la Figura 6.13.

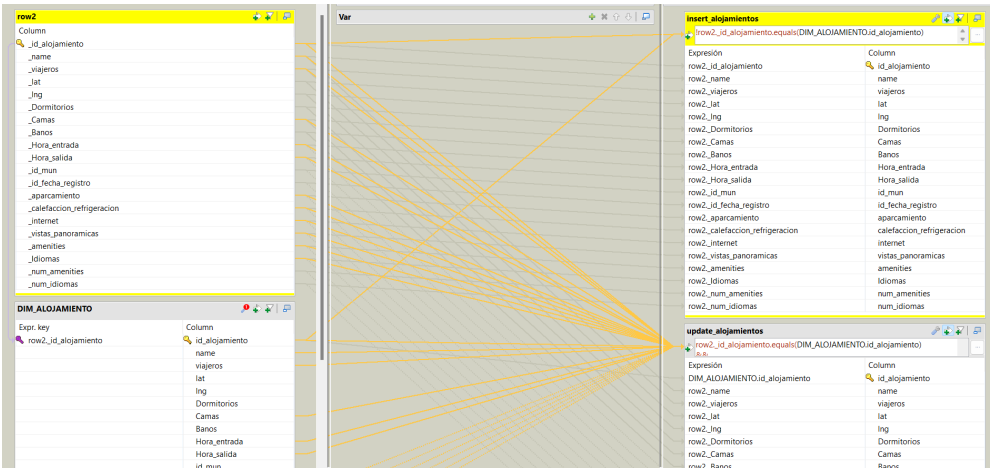

Figura 6.11: tMap Dimensión Alojamientos

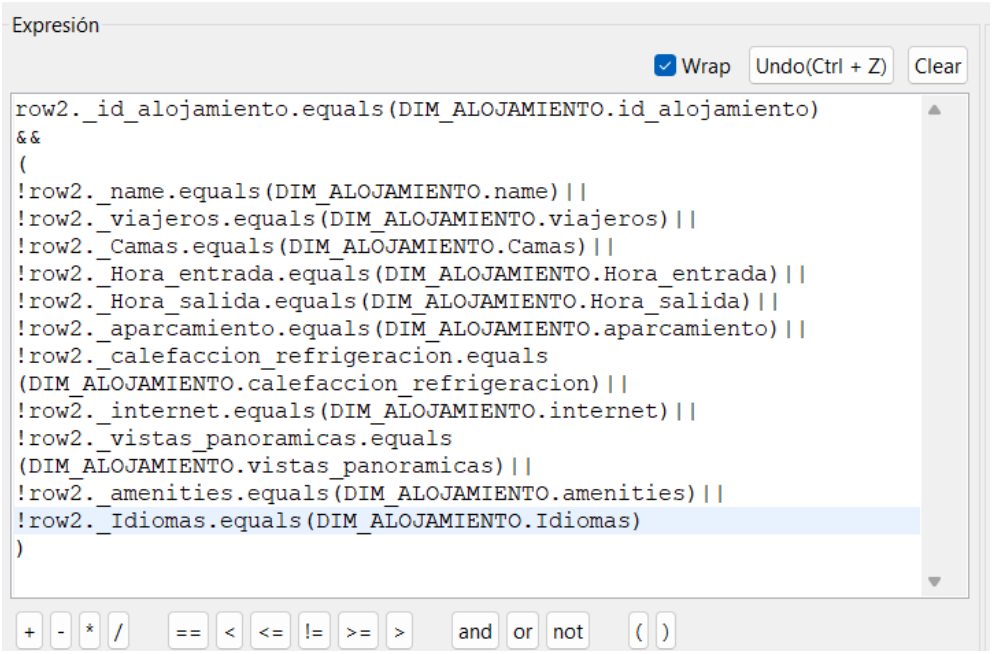

Figura 6.12: Expresión tMap-Update Alojamientos

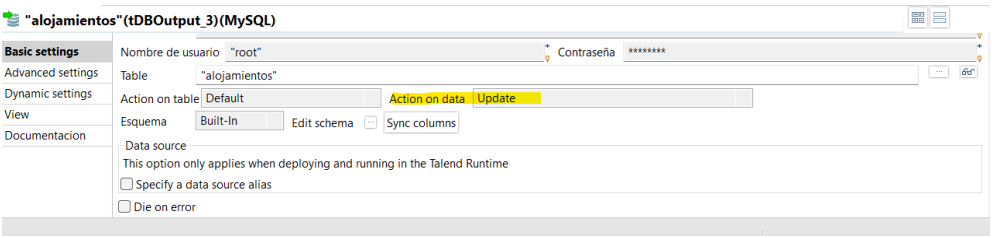

Figura 6.13: Marcar valor de Update Alojamientos

Para comprobar el correcto funcionamiento tanto de la función de insertar nuevos alojamientos, como de actualizar, se modificaron los cuatro primeros alojamientos y se insertaron dos alojamientos nuevos (una vez hecha la comprobación se volvió a los datos originales). En la Figura 6.14 puede observarse que la carga se realizó de la manera esperada, comprobándose después que la información era correcta también desde base de datos. En la Figura 6.15 se muestra cómo se actualizaron las horas de entrada de los alojamientos con identificador '0', '2' y '3', y el número de camas y viajeros en el alojamiento '1'. Por otro lado, en la Figura 6.16 se puede ver la inserción de los dos nuevos alojamientos.

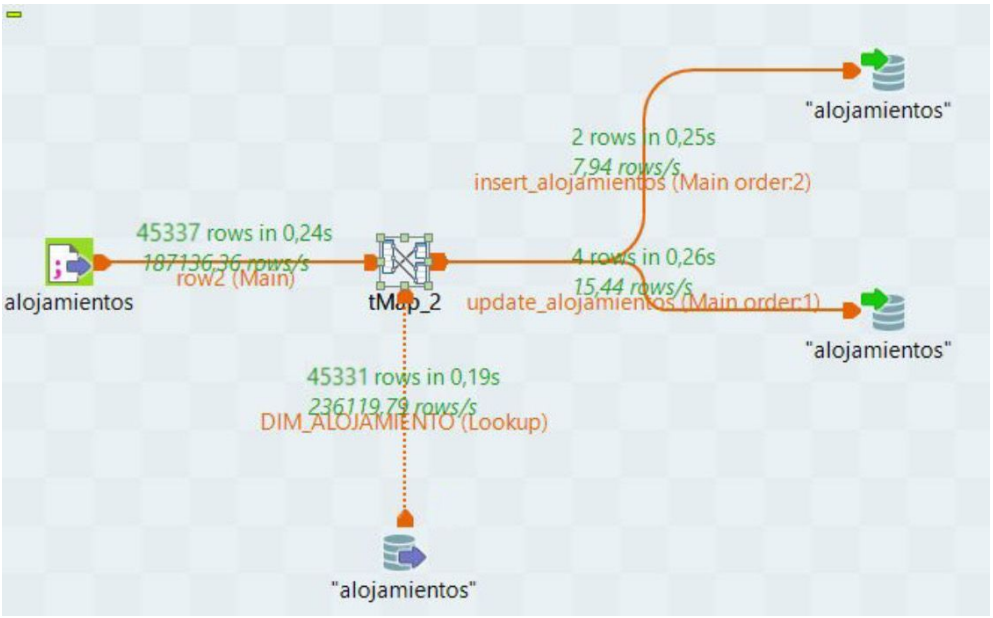

Figura 6.14: Comprobación funcionamiento tMap en la Dimensión Alojamientos

	id_alojamiento	name	viajeros	lat	lng	Dormitorios	Camas	Baños	Hora_entrada
1	0	Apartamento de lujo 1ª linea de Playa Finisterre	6	42,91	-9,26	2	4	2	12:00:00
2	1	Loft Finisterre Centro	3	42,9	-9,26	1	2	5	15:00:00
3	2	Casa Real 43. Casa marinera con vistas al mar	2	42,9	-9,26	1	1	1	12:00:00
4	3	Finisterre Centro	4	42,9	-9,26	2	3	1	13:00:00
5	4	Casita marinera con vistas al mar y a la montaña	4	42,9	-9,26	1	1	5	15:00:00
6	5	APARTAMENTO ESTUDIO O ALMACEN	4	42,91	-9,26	1	1	5	15:00:00
7	6	CASA ENTERA TALON , playa de Nemiña	4	43,01	-9,27	2	[NULL]	1	[NULL]
8	7	ACCESO PLAYA- PISCINA- PADEL. FISTERRA.GrupoL3. 1C	6	42,91	-9,26	3	6	2	11:00:00

	id_alojamiento	name	viajeros	lat	lng	Dormitorios	Camas	Baños	Hora_entrada
1	0	Apartamento de lujo 1ª linea de Playa Finisterre	6	42,91	-9,26	2	4	2	16:00:00
2	1	Loft Finisterre Centro	2	42,9	-9,26	1	1	5	15:00:00
3	2	Casa Real 43. Casa marinera con vistas al mar	2	42,9	-9,26	1	1	1	15:00:00
4	3	Finisterre Centro	4	42,9	-9,26	2	3	1	15:00:00
5	4	Casita marinera con vistas al mar y a la montaña	4	42,9	-9,26	1	1	5	15:00:00
6	5	APARTAMENTO ESTUDIO O ALMACEN	4	42,91	-9,26	1	1	5	15:00:00
7	6	CASA ENTERA TALON , playa de Nemiña	4	43,01	-9,27	2	[NULL]	1	[NULL]
8	7	ACCESO PLAYA- PISCINA- PADEL. FISTERRA.GrupoL3. 1C	6	42,91	-9,26	3	6	2	11:00:00

Figura 6.15: Comprobación de actualización en la Dimensión Alojamientos

	id_alojamiento	name	viajeros	lat	lng	Dormitorios	Camas	Baños	Hora_entrada
1	110.078	FIC2	2	43,33	-8,41	1	1	1	13:00:00
2	110.077	FIC1	4	43,33	-8,41	2	2	1	12:00:00
3	110.076	apartamento tranquilo y espacioso 3 dormitorios y 2 planta baja r	6	35,14	-2,91	3	4	2	[NULL]
4	110.075	Apartamento confort Nador Al Mataar	6	35,17	-2,93	3	5	1	14:00:00
5	110.074	Alquila un bonito apartamento en el centro de Nador.	10	35,16	-2,92	3	10	1	[NULL]

Figura 6.16: Comprobación de inserción de nuevos registros en la Dimensión Alojamientos

Hecho Gasto Total

Una vez finalizada la carga de las dimensiones se procedió con las tablas de hechos.

La tabla de hechos de 'Gasto Total' almacena los datos mensuales del gasto total en turismo; por lo tanto, estos datos no necesitarán de actualizaciones, pero sí de nuevas inserciones. Como en casos anteriores se incorpora el componente 'tMap', en este caso con dos archivos de origen (el fichero CSV con los datos y la conexión a la tabla de hechos correspondiente de

la base de datos) y una salida (la tabla de hechos de la base de datos).

Para tratar la inserción sin que cada vez que se inserten nuevos datos se reinserten todos dando lugar a duplicados se ha procedido de forma similar a como se ha hecho en los procesos de inserción utilizados con anterioridad. El mapeo realizado se puede ver en la Figura 6.17.

Figura 6.17: tMap Hecho Gasto Total

La Figura 6.18 muestra la carga inicial de los datos de esta tabla de hechos; mientras que en la Figura 6.19 se puede ver cómo una vez cargados los datos, si no hay información nueva, no se inserta.

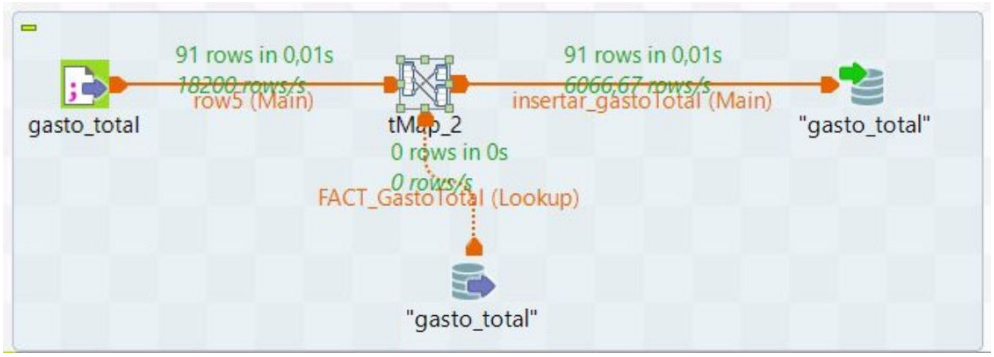

Figura 6.18: Carga Hecho Gasto Total

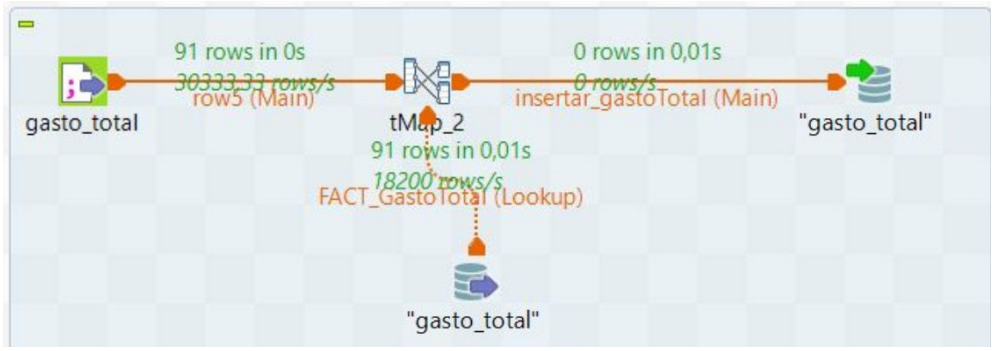

Figura 6.19: Insertar datos duplicados en el Hecho Gasto Total

Hecho INE

En el caso del hecho que contiene todos los demás datos extraídos del INE ocurre lo mismo; podrán darse nuevas inserciones pero no será necesario actualizar datos antiguos, por tanto, se procederá de igual manera.

En la Figura 6.20 se ve el mapeo de los datos. En este caso, la inserción o no de un dato se decide haciendo uso no solo del identificador de fecha (como ocurría en la tabla de hechos de 'Gasto total') sino también con el identificador de la Comunidad Autónoma; ya que para una misma fecha existen diferentes registros en las diferentes regiones. En la Figura 6.21 se aprecia la carga inicial; la comprobación de que no se insertan datos ya insertados se observa en la Figura 6.22.

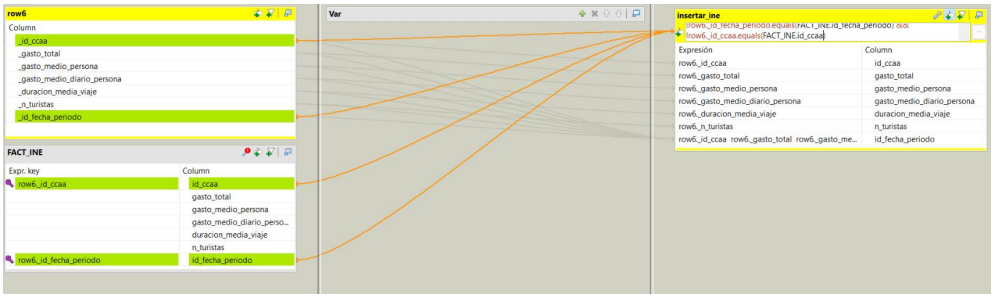

Figura 6.20: tMap Hecho INE

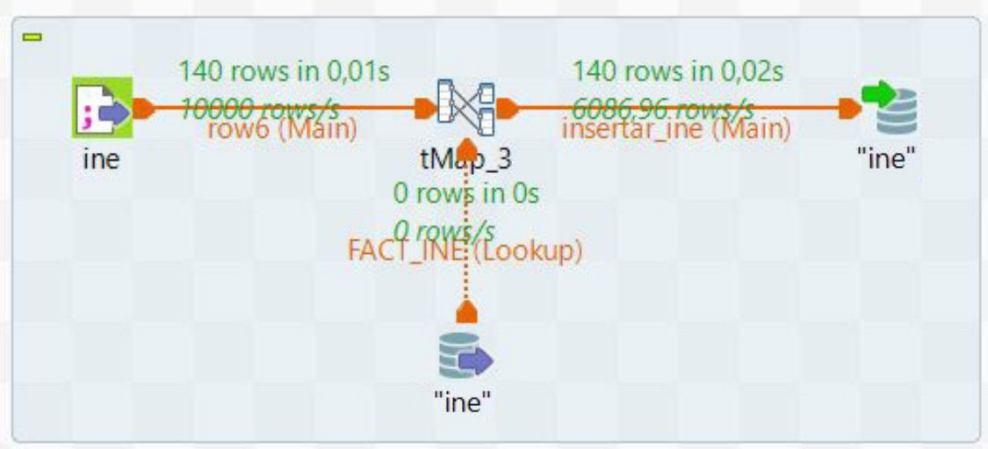

Figura 6.21: Carga Hecho INE

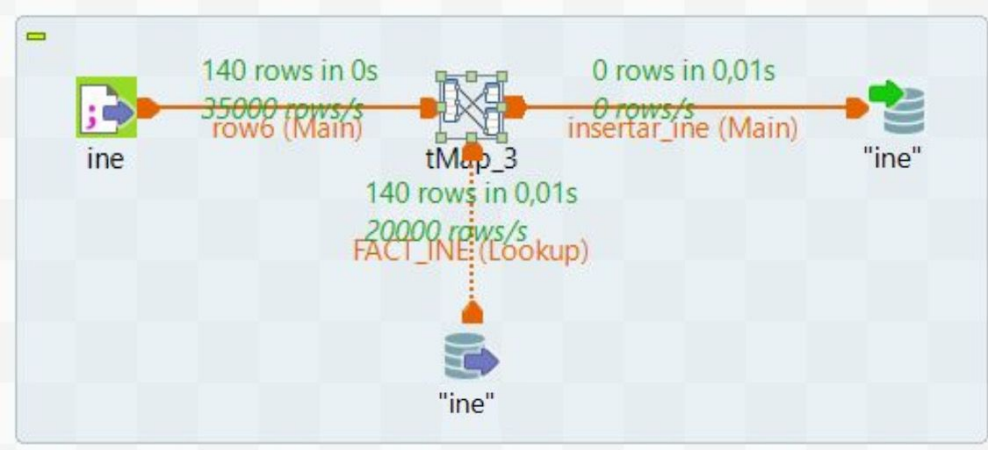

Figura 6.22: Insertar datos duplicados en el Hecho INE

Hecho Alquiler

Finalmente, en la tabla de hechos de 'Alquiler', se siguen pasos similares a los indicados para la dimensión de los alojamientos. En este caso se necesitan de nuevo dos salidas, una para inserciones y otra para actualizaciones, con la diferencia de que en este caso cada actualización se insertará como una fila nueva y no como una actualización de la fila. Esta decisión se ha tomado para poder ver la evolución de las diferentes medidas en el tiempo.

En la Figura 6.23 se muestra el mapeo de los datos. Aunque en este hecho, en ambos casos (inserción y actualización) se insertará una nueva fila con el dato nuevo o el dato modificado, se ha optado por tener dos conexiones de salida. La inserción se realizará cuando el identificador de alojamiento aún no se encuentre en la base de datos, como sucedía con la dimensión 'Alojamiento', mientras que la actualización se producirá cuando este identificador sí exista pero haya algún cambio en alguna de las columnas que figuran en la expresión de la Figura 6.24.

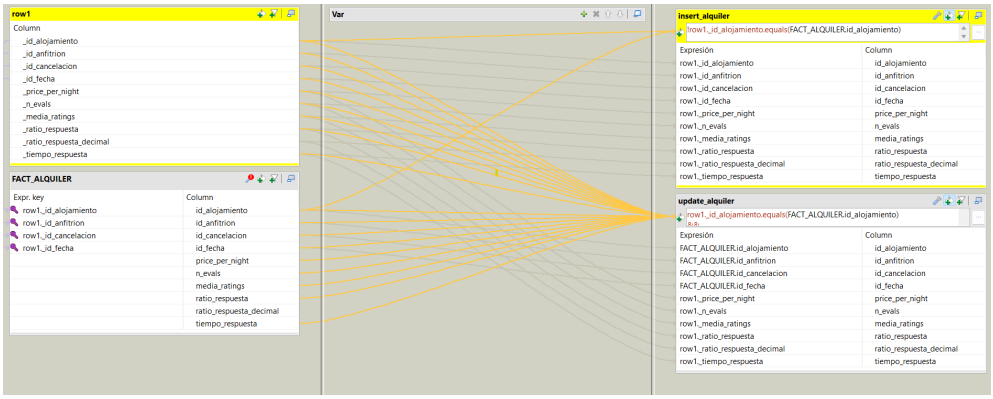

Figura 6.23: tMap Hecho Alquiler

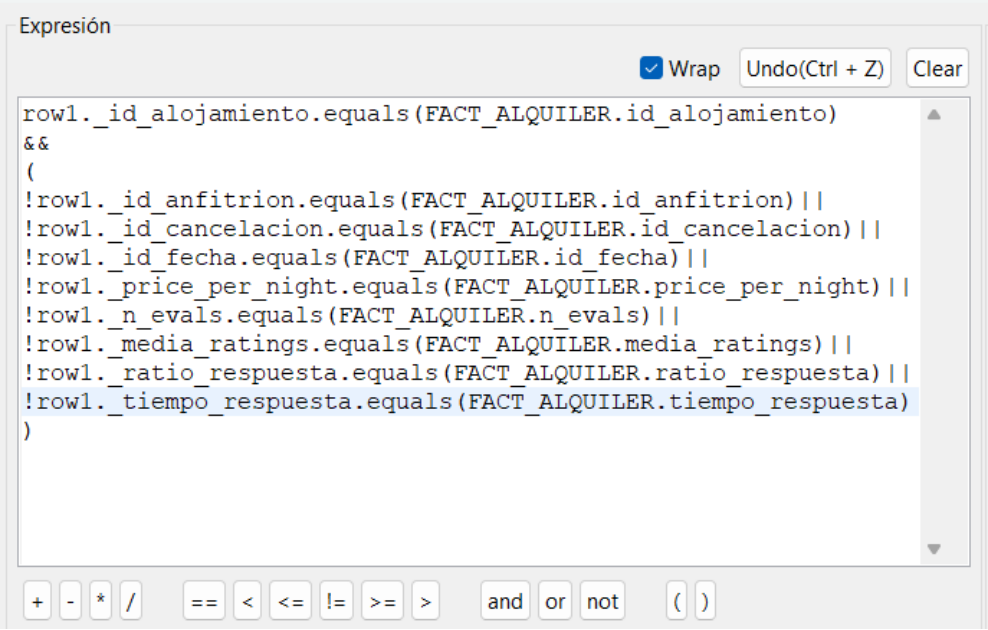

Figura 6.24: Expresión tMap-Update Alquiler

Tras definir las especificaciones del componente de mapeo se procede con la carga inicial de los datos (Figura 6.25) y, a continuación, con una comprobación similar a la hecha en la dimensión de los alojamientos. Se hacen en este caso también modificación en cuatro de las filas del CSV y se insertan dos nuevos alojamientos para comprobar que la carga se realiza de la forma esperada. En la Figura 6.26 se ve cómo, en un principio, se produjo la actualización de 4 filas y la inserción de otras 2; las figuras, Figura 6.27 y Figura 6.28, muestran, mediante consulta a base de datos, cómo, efectivamente, el resultado de dichas operaciones fue el deseado.

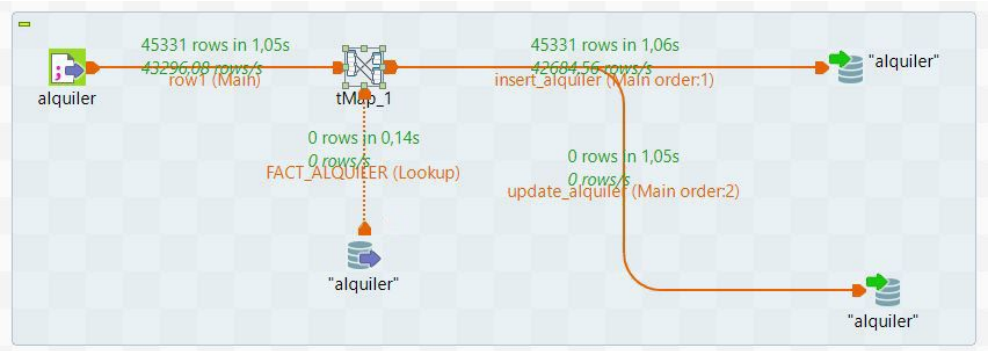

Figura 6.25: Carga Hecho Alquiler

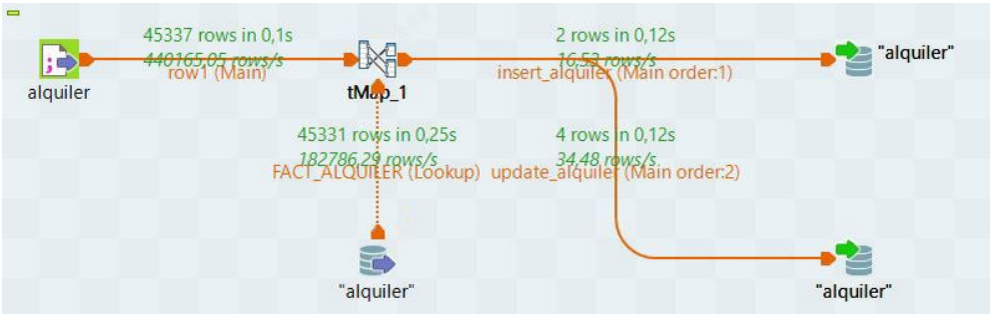

Figura 6.26: Comprobación funcionamiento tMap en el Hecho Alquiler

	id_alojamiento	id_anfitrion	id_cancelacion	id_fecha	price_per_night	n_evals	media_ratings	ratio_respuesta	ratio_respu
1	0	0	2	0	59	202	4,82	100%	
2	0	1	3	3	59	212	4,82	100%	
3	1	0	2	0	[NULL]	120	4,72	100%	
4	1	1	3	3	[NULL]	323	4,72	100%	
5	2	0	2	0	80	22	5	100%	
6	2	0	1	3	80	26	5	100%	
7	3	0	1	0	[NULL]	120	4,85	100%	
8	3	1	1	3	[NULL]	120	4,95	100%	
9	4	0	2	0	65	69	4,96	100%	
10	5	0	1	0	89	292	4,93	100%	

Figura 6.27: Comprobación de actualización en el Hecho Alquiler

alquiler | Enter a SQL expression to filter results (use Ctrl+Space)

Grilla | Texto

	id_alojamiento	id_anfitrion	id_cancelacion	id_fecha	price_per_night	n_evals	media_ratings	ratio_respuesta	ratio_respuesta_d
1	110.078	1	0	3	33	114	4,88	100%	
2	110.077	1	0	3	58	2	[NULL]	77%	
3	110.076	1	0	0	54	2	[NULL]	67%	
4	110.075	1	3	0	50	3	[NULL]	0%	

Figura 6.28: Comprobación de inserción de nuevos registros en el Hecho Alquiler

Capítulo 7

Explotación del almacén de datos

Una vez creado el Data Warehouse, el objetivo de este capítulo es la explotación de los datos, estructurados y almacenados en base de datos, en base a las necesidades de análisis establecidas inicialmente. Como se podrá apreciar a continuación, los siguientes apartados pretenden cubrir, mediante la creación de distintos *dashboards*, cada uno de los objetivos de análisis planteados en la Sección 5.1.1. El objetivo 'Gastos y precios' no se verá reflejado en un *dashboard* en particular, sino que será tratado de forma global en todos ellos. En cada *dashboard* la información podrá ser filtrada por años para así poder contrastar diferencias y ver su evolución a lo largo del tiempo.

Para poder elaborar este análisis visual de una manera óptima se ha decidido hacer uso de la herramienta PowerBI, descrita en la Sección 3.1.3.

A mayores, se presenta también, a modo de muestra, la posibilidad de desarrollar modelos a partir de nuestro conjunto de datos para poder realizar predicciones.

7.1 Dashboard 1. Visión general

7.1.1 Finalidad

Este primer *dashboard*, Figura 7.1, sirve para dar una perspectiva general global de la evolución del turismo a lo largo de los años (y meses) en España. A partir del análisis de alojamientos nuevos registrados y del gasto total destinado a turismo se introduce al usuario en materia.

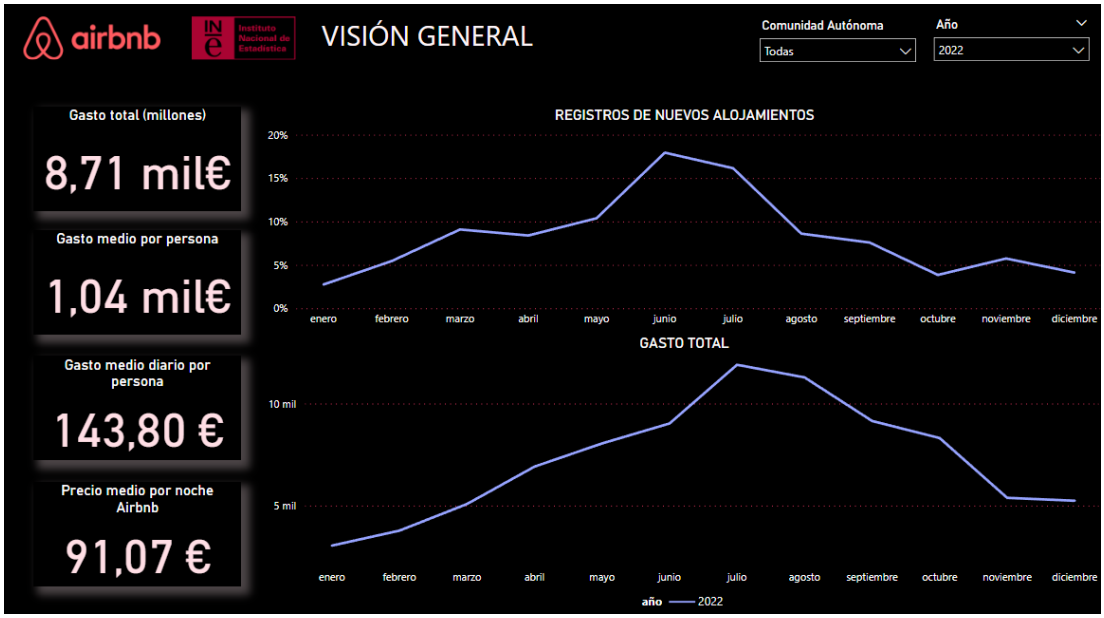

Figura 7.1: Dashboard 1 - Evolución del turismo a lo largo del tiempo

7.1.2 Descripción

En la zona central de este *dashboard* se pueden observar dos gráficos en los que se representa la evolución por meses del gasto total en turismo y de los registros de nuevos apartamentos en Airbnb. En el lateral izquierdo se incluyen cuatro tarjetas informativas que muestran diferentes valores referentes a los gastos y el precio de los alojamientos.

Toda la información de esta visualización se puede ir seleccionando y filtrando por año. Así, el usuario puede ver y comprobar las diferencias en el tiempo, conocer qué años ha habido mayor ganancia en el sector turismo y qué años se ha resentido. Además, se ha incluido el filtro por Comunidad Autónoma para así poder observar las diferencias entre cada una de ellas.

7.1.3 Análisis

En los gráficos se puede ver claramente cómo va aumentando el número de nuevos apartamentos en la plataforma en los meses previos a verano, momento en el cual también se puede observar que es cuando más gasto en turismo realiza la población y a partir del cual luego comienza a descender de nuevo.

Se puede comprobar también cómo los diferentes anfitriones tienen bastante claro cuándo deben poner sus alojamientos a disposición en la aplicación, siguiendo la tendencia de aumento del gasto en el sector. Desde la posición de huésped también se puede hacer apunte de esta curiosa situación. Si la intención inicial es hacer una reserva desde la plataforma de

Airbnb para lo meses de verano, aunque muchos de los viajeros quisieran dejar las habitaciones reservadas con meses de antelación, no se deben apresurar ni agobiar si al planificar estas vacaciones no encuentran alojamiento porque, según se acerquen más las fechas estivales, aparecerán nuevas ofertas.

Finalmente, haciendo uso de este *dashboard*, y aplicando un filtro por el año 2020 (Figura 7.2), se puede comprobar la repercusión que tuvo el Covid-19 en este sector: cómo se ve reducido sustancialmente el gasto total y cómo hay un gran descenso en los primeros meses de la enfermedad tanto en los registros de la aplicación como en el mencionado gasto.

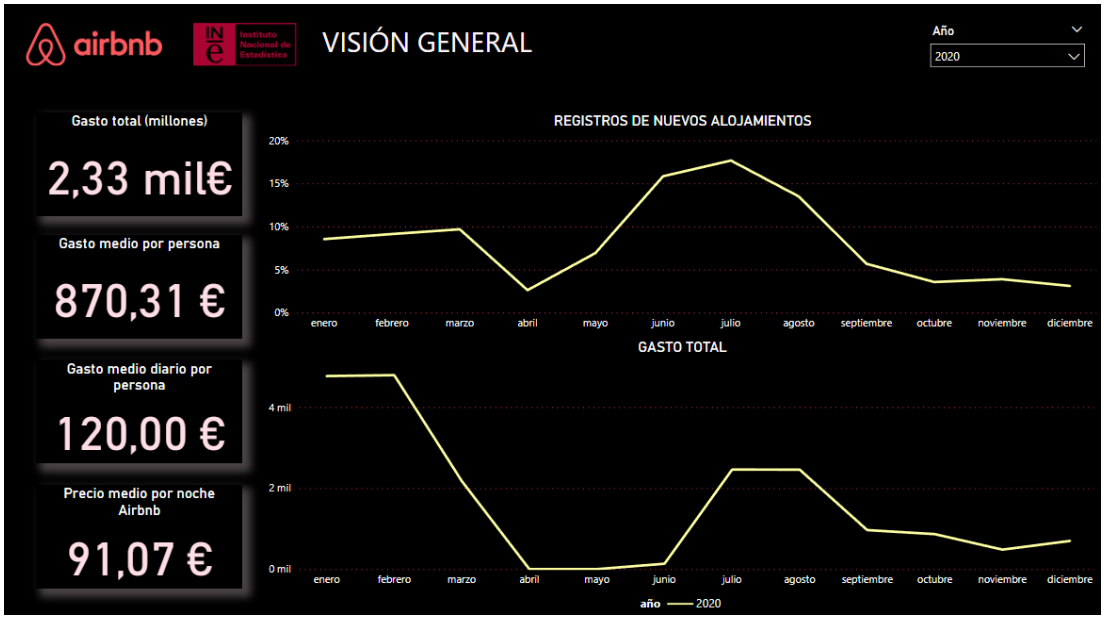

Figura 7.2: Dashboard 1 - Covid-19

7.2 Dashboard 2. Comparativa INE y Airbnb

7.2.1 Finalidad

Este segundo *dashboard*, Figura 7.3, se concibe como ayuda para la toma de decisiones en el sector turismo. En él se pueden encontrar diferentes correlaciones entre características, realizando un estudio de los datos del INE junto con los de Airbnb, para determinar indicadores que faciliten, como propietarios, tomar decisiones en la oferta de alojamientos, o incluso como analizadores del sector, para contemplar futuros cambios que puedan potenciar el turismo.

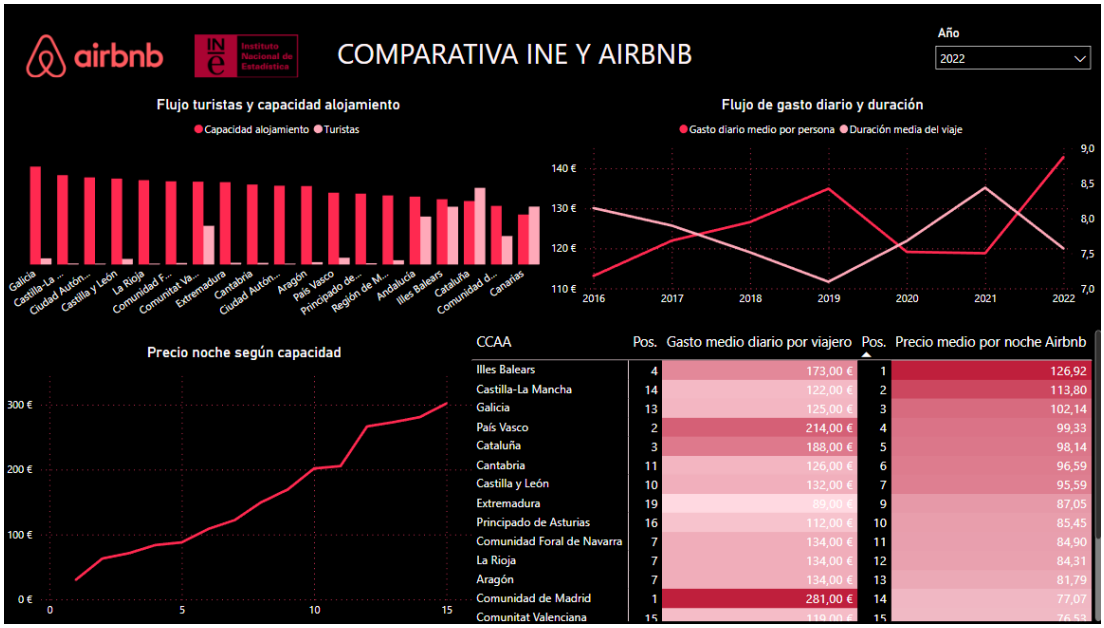

Figura 7.3: Dashboard 2 - Comparativa INE vs. Airbnb

7.2.2 Descripción y análisis

En la primera visualización, el gráfico de barras que se localiza en la parte superior izquierda, muestra la distribución de turistas por Comunidad Autónoma en comparación con la capacidad de los alojamientos de Airbnb. Con ello se puede comprobar cómo las comunidades autónomas con más turismo, exceptuando Valencia, son las que menos capacidad tienen. Esto es un indicativo de que hay una mayor demanda de apartamentos pequeños y de que los viajes, por tanto, suelen ser de grupos pequeños y parejas. En estos casos se podría tomar nota como anfitriones y dar mayor oferta de alojamientos con capacidades más bajas, o incluso como región, intentar aumentar el número de los mismos para estudiar si de esta manera también aumenta el de viajeros.

En la zona superior derecha vemos como se relacionan el flujo de gasto diario con la duración del viaje. Se puede observar cómo siguen tendencias opuestas, es decir, cuando hay mucho gasto diario el viaje suele ser más corto. Esto es un comportamiento lógico, que indica que hay un presupuesto medio, el cual se puede emplear en más días y menos gasto o en más gasto y, por tanto, menos días de viaje. Teniendo esto en cuenta, si se quisiera ampliar la duración de los viajes en un determinado lugar se podría desarrollar un modelo de negocio alternativo para ofrecer alojamientos más baratos y que los turistas disfrutasen de más días de vacaciones. Una alternativa podría pasar por aumentar el número de alojamientos compartidos o implementar opciones de servicios de bajo costo.

Por otro lado, en la zona inferior izquierda, se muestra una gráfica de cómo el precio por

noche del apartamento aumenta según su capacidad. Estas dos variables están claramente correlacionadas.

Por último, se incluye una tabla en la cual se compara el gasto diario por viajero con el gasto por noche en Airbnb. Se puede ver cómo no hay una clara relación entre ambos; aunque en conjunto, las Islas Baleares, País Vasco y Cataluña se posicionan como las comunidades con los precios más altos, teniendo en cuenta sendas variables. En estas regiones resultaría interesante buscar la forma de contemplar los servicios de bajo costo anteriormente mencionados y comprobar el incremento de noches en las mismas. Para comprobar las repercusiones de este cambio y otros que se puedan implementar se cuenta con el filtro por año.

Además, algo curioso a destacar es lo que ocurre con la Comunidad de Madrid, en la que el gasto por viajero es el más elevado mientras que en Airbnb no destaca por sus precios. Esto podría deberse a que en esta ciudad en los últimos dos años se ha incrementado el alquiler por habitación y no por apartamento, lo cual abarata mucho los gastos por noche.

7.3 Dashboard 3. Evaluaciones

7.3.1 Finalidad

Como indica el título y cómo se puede apreciar en la Figura 7.4, este *dashboard* se centra en el análisis de evaluaciones ofrecidas por los usuarios, con el objetivo de comprobar cómo estas afectan a un alojamiento junto con sus diferentes servicios. Resulta muy interesante saber como anfitrión hasta qué punto se debe cuidar y prestar atención a esta valoración y qué se podría hacer para que ésta mejore.

7.3.2 Descripción

En este *dashboard* contamos con un mayor número de filtros debido a que se han considerado varias características de interés en lo que a evaluaciones respecta. Los filtros que se han incluído son los siguientes:

- Número de evaluaciones: La información de este *dashboard* puede ser filtrada por el número de evaluaciones ya que, como es lógico, no es lo mismo 5 estrellas de 1 valoración que esas mismas 5 estrellas de 100 valoraciones.

- Año: Como se viene mostrando hasta el momento, aquí también se ha incluído un filtro por año. En este momento no es un filtro que pueda ser explotado como tal, pues toda la información referente a Airbnb ha sido recogida este año. Sin embargo, para un futuro, es un filtro de gran interés.

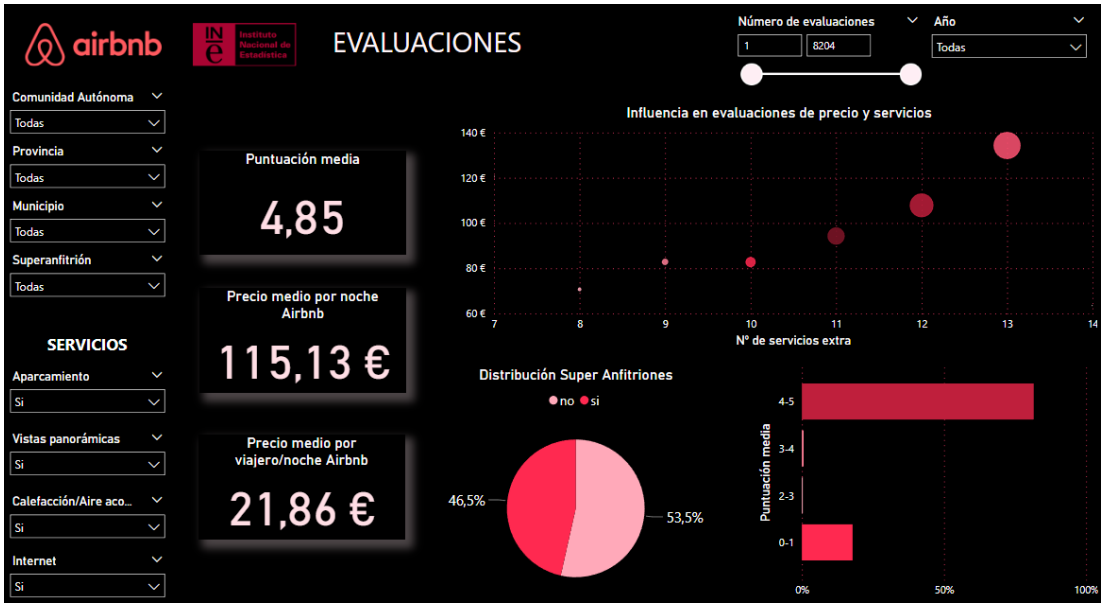

Figura 7.4: Dashboard 3 - Evaluaciones

- Comunidad Autónoma, Provincia y Municipio: En este caso se han incluido estos tres campos, ya que los datos recogidos por Airbnb cuentan con la precisión por municipios y no solo por Comunidad Autónoma, permitiéndonos así poder ver las diferencias entre municipios, si fuese necesario.
- Super Anfitrión: La característica del tipo de anfitrión es de relevancia a la hora de llevar a cabo un estudio de las evaluaciones ya que, como se analizará a continuación, los Super Anfitriones cuentan con mejores valoraciones.
- Servicios: Además, se ha tomado la decisión de filtrar por los cuatro servicios considerados prioritarios; de esta manera se podrá ver su influencia en la satisfacción de los usuarios durante su estancia.

De igual forma, este *dashboard* cuenta también con tres tarjetas informativas con valores relativos a la evaluación media, el precio medio por noche y el precio medio por viajero y noche. Este último se obtiene de la división del precio por noche de los alojamientos entre su capacidad. A mayores, se incluyen tres visualizaciones relativas a las evaluaciones, el precio, el número de servicios y los anfitriones.

7.3.3 Análisis

En relación a las tarjetas informativas, puede apreciarse cómo los tres valores aumentan según se añaden servicios a un apartamento; por tanto, si un anfitrión busca poder subir el precio de un alojamiento u obtener mejores evaluaciones debe ofrecer algunos de ellos.

Además, en el caso de las evaluaciones, un dato relevante es el de los Súper Anfitriones. Estos son anfitriones calificados por Airbnb como tal debido a su buen servicio y cumplimiento de una serie de requisitos. Su porcentaje también aumenta si se contemplan los diferentes servicios que se pueden filtrar.

Si comparamos la Figura 7.5 con la Figura 7.4 podemos ver lo señalado en los dos puntos anteriores.

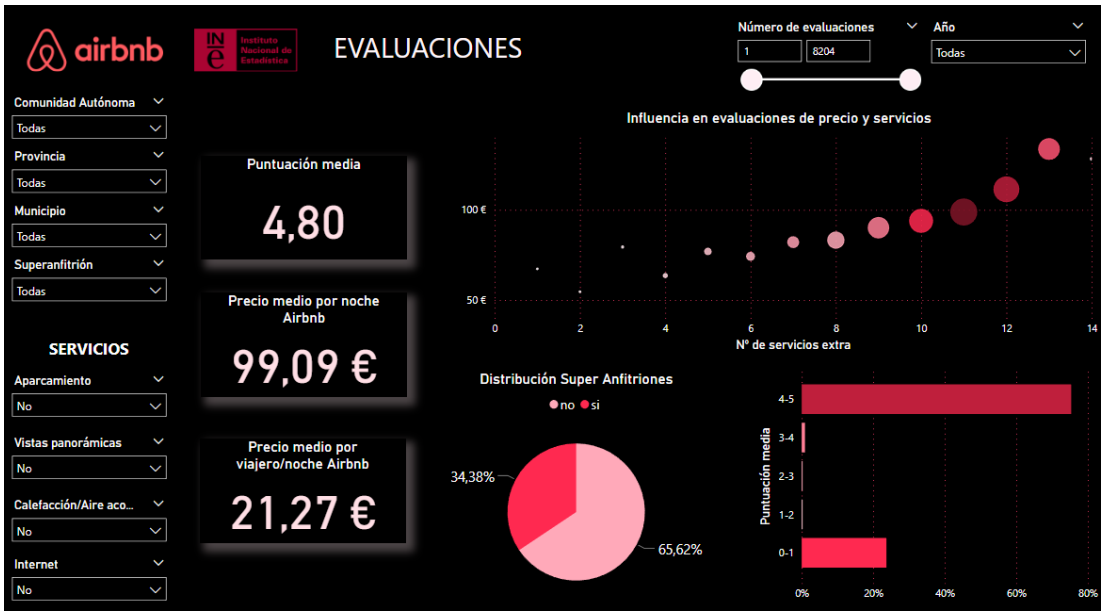

Figura 7.5: Dashboard 3 - Evaluaciones filtrando sin los servicios

Por otro lado, en este *dashboard* puede verse una gráfica que muestra la influencia del precio y los servicios en las evaluaciones. Se puede comprobar cómo efectivamente se confirma la conclusión anterior; cuantos más servicios tiene un alojamiento más aumenta su precio y su evaluación media.

Finalmente, se incluye una distribución de la puntuación media y cómo los usuarios muestran su posición cuando están muy satisfechos o muy insatisfechos. Además, si se selecciona el rango 0-1 se ve cómo el número de Súper Anfitriones disminuye, lo cual ocurre a la inversa si se seleccionan las evaluaciones en el rango 4-5.

En conclusión, si un propietario quiere sacarle un buen rendimiento y partido a su alojamiento debe ofrecer la mayor cantidad de servicios. Una vez hecho esto, las buenas calificaciones empezarán a llegar y también el distintivo de Super Anfitrión, el cual dará una seguridad extra al huésped. Una vez obtenga esto, él mismo podrá subir el precio de su apartamento y exprimirlo al máximo.

7.4 Dashboard 4. Comparativa Galicia y España

7.4.1 Finalidad

Este *dashboard*, Figura 7.6, busca mostrar una imagen comparativa de Galicia frente al resto de España, y con ello servir de ayuda para la implementación de medidas que favorezcan un posible incremento del sector turístico en nuestra comunidad.

Figura 7.6: Dashboard 4 - Comparativa Galicia vs.España

7.4.2 Descripción

Al igual que en otros casos, aquí también se cuenta con los filtros por año, para poder explorar las ya mencionadas evoluciones de los datos en el tiempo, y por Comunidad Autónoma, para que la comparativa pueda realizarse no solo con valores de España sino de cualquier otra comunidad.

En la zona superior se muestran dos gráficos de barras y líneas. En el primero de ellos se pueden ver los flujos de duración del viaje y gasto diario, que, como ya se comentó previamente, seguían un comportamiento opuesto. Ahora se puede ver cómo Galicia destaca por ser un destino turístico al cual se va a pasar un mayor número de días al requerir un gasto más bajo que la media.

En el segundo gráfico se muestra el flujo de turistas y el gasto total, variables ambas directamente correlacionadas. En este caso se puede ver cómo Galicia cuenta con un número bajo de turistas y por tanto el gasto total en turismo también es inferior a la media, ya que

aunque los viajes sean más largos, como bien se ha dicho, también el gasto diario es menor, por lo que el gasto por persona será el mismo que en otras comunidades.

Finalmente, se incluyen diferentes tarjetas comparativas que muestran, para distintas características, los valores de España (o la Comunidad Autónoma que se seleccione en los filtros) y el valor de Galicia. En cuanto al precio por persona y noche del alojamiento de Airbnb se ve cómo efectivamente el valor para nuestra comunidad es algo menor que la media en España, lo cual corrobora lo indicado respecto al gasto diario. También se aprecia cómo la capacidad media del alojamiento es algo más elevada, lo cual es indicativo de que suelen ser grupos de amigos o familias los turistas que vienen a Galicia. En los indicadores de la derecha (zona superior) se muestra el gasto total, el cual es efectivamente inferior. Además, también se observa un dato curioso en el precio por noche en Airbnb (zona inferior), que es más alto que la media del resto del país, pero como se deduce, esto es debido a que la capacidad en nuestra comunidad suele ser también mayor.

7.4.3 Análisis

De acuerdo con lo observado en las gráficas superiores, un método para aumentar el gasto total en turismo en Galicia no debiera ser aumentar los precios, ya que apunta a que muchos de los turistas optan por este destino y sus largas estancias por el menor gasto que supone. En su lugar, el sector turístico debería centrarse en incrementar el número de turistas aprovechando esa ventaja económica. Esto podría lograrse mediante la promoción activa de Galicia como un destino asequible y atractivo. Al aumentar el número de visitantes, es probable que se genere un aumento en el gasto total en turismo, lo que beneficiaría tanto a la industria turística como a la economía local, en general.

En ambos gráficos, y de forma más general, se puede corroborar lo observado anteriormente con respecto a la pandemia. Se puede ver el descenso que sufre 2020 con respecto a los demás años y cómo a partir de éste se da una recuperación a valores previos al Covid-19.

Analizados los valores de las tarjetas, otra estrategia de negocio que podría ser interesante a seguir, como ya se ha mencionado antes, sería aumentar el número de alojamientos de baja capacidad en Galicia. Este enfoque se basa en la idea de que Galicia podría ser un destino muy atractivo para aquellos que deseen pasar unas vacaciones en pareja o en solitario, buscando la tranquilidad y la serenidad que esta región ofrece. Se podría promocionar este tipo de alojamientos a través de estrategias de márketing dirigidas específicamente a parejas o viajeros individuales, generando un interés adicional y atrayendo a un segmento de mercado más específico.

Resumiendo, el sector turístico de Galicia podría seguir una estrategia que diese a conocer los precios asequibles con los que cuenta esta comunidad y realizase un aumento de alojamientos de capacidades más bajas.

7.5 Modelado y predicción

Para poder mostrar, de manera sencilla, la capacidad de realizar también predicciones sobre los datos extraídos y almacenados, se ha optado por predecir el precio por noche de los alojamientos, en función de sus características. Esta predicción resulta de interés tanto para anfitriones, para así poder mantener un precio competitivo en sus alojamientos, como para turistas, quienes pueden decidir el precio que están dispuestos a pagar por un apartamento que cumpla con sus necesidades.

7.5.1 Modelado

Los datos utilizados para la predicción se obtuvieron mediante una consulta a base de datos con la que se recuperó información relevante sobre los alojamientos, incluyendo el precio por noche, el número de viajeros, el número de dormitorios, el número de camas, el número de baños, el número de extras y el municipio al que pertenecen.

Antes de construir los modelos de predicción, se dividió el conjunto de datos en conjuntos de entrenamiento y prueba, utilizando un 70% de los datos para entrenamiento y el 30% restante para pruebas.

Se exploraron varios modelos de aprendizaje automático para predecir el precio de los alojamientos. Los modelos utilizados fueron Bosques Aleatorios (*Random Forest*), *XGBoost* y Regresión Lineal con Regularización (*Ridge*). Para todos ellos se fueron haciendo pruebas con diferentes hiperparámetros hasta encontrar los óptimos.

Una vez obtenidos los diferentes resultados de cada uno de los modelos de predicción (Figura 7.7), se compararon utilizando métricas de rendimiento como el RMSE (Error Cuadrático Medio) y el R-squared (Coeficiente de Determinación). El modelo que obtuvo las mejores métricas fue el de Random Forest, siendo así seleccionado como modelo final.

```
> # Obtener la media del RMSE y del Rsquared
> rmse <- summary_results$statistics$RMSE[, "Mean"]
> rmse
   Random_Forest          XGBoost Ridge_Regression
        11.94656         25.52078         34.32275
> rsquared <- summary_results$statistics$Rsquared[, "Me
> rsquared
   Random_Forest          XGBoost Ridge_Regression
       0.8596067        0.6753075        0.5980783
```

Figura 7.7: RMSE y R-squared de los diferentes modelos

7.5.2 Predicciones

Una vez entrenado y evaluado el modelo final es posible realizar ya predicciones sobre nuevos datos. Se proporciona un ejemplo de cómo utilizar el modelo entrenado para predecir el precio de un nuevo alojamiento en función de sus características en la Figura 7.8.

```
> #Predecimos
> # Crear un nuevo conjunto de datos con las características de predicción
> new_data <- data.frame(
+   viajeros = 2,
+   Dormitorios = 1,
+   Banos = 1,
+   Camas = 1,
+   num_amenities = 3,
+   id_ccaa = 3 #Asturias
+ )
> # Realizar una predicción
> predicted_price <- predict(model_rf, newdata = new_data)
> predicted_price
       1
48.30418
>
```

Figura 7.8: Ejemplo de predicción sobre alojamiento

El modelo final seleccionado representa una herramienta útil para estimar el precio por noche de los alojamientos en función de sus características, lo que, como ya se ha indicado, puede ser muy valioso para los propietarios de alojamientos, pero también para los propios viajeros que buscan alojamiento en Airbnb.

Capítulo 8

Conclusiones

Al igual que en la introducción de este proyecto, cabe recordar aquí la importancia que tiene España como destino turístico y, por tanto, la relevancia de este sector para la economía de nuestro país. Gracias a este trabajo se ha podido profundizar en el estudio de este sector y evidenciar algunas de sus características.

A lo largo de este proyecto se han realizado importantes procesos y desarrollos. Se ha comprendido e implementado un completo proceso ETL, extrayendo datos de las diferentes fuentes, procesándolos y cargándolos en un almacén de datos. Este proceso a su vez ha necesitado de la utilización y comprensión del funcionamiento de un Data Warehouse, así como poder llevar a cabo un diseño que cumpliese con todas las necesidades requeridas por este trabajo.

Cabe mencionar que el proceso de extracción de datos ha sido un proceso bastante laborioso, tanto computacionalmente como a la hora de diseñar código, necesitándose de varias extracciones de prueba hasta obtener la adecuada.

Por su parte, el diseño, transformación y carga de datos se puede afirmar también que son procesos especialmente delicados, en los cuales se debe prestar atención a los detalles, ya que mínimos fallos pueden ocasionar errores con importantes repercusiones en el modelo de datos.

Completado el proceso ETL, se ha procedido a la explotación de los datos obtenidos realizando diversos análisis, mediante la creación de *dashboards* interactivos y visualizaciones, que permitiesen extraer información de interés, a partir de grandes volúmenes de datos, de manera sencilla y óptima. Con ello, se han podido cubrir las necesidades fijadas en el momento de establecer los objetivos de este proyecto, poniendo en relieve, entre otros, la relación de los datos de alojamientos y el turismo en España, precios, gastos, etc. Además, se ha mostrado la capacidad de poder realizar predicciones en base al conjunto de datos almacenado; de interés también para una futura ampliación de los trabajos realizados.

8.1 Trabajo futuro

Tras cumplir los objetivos fijados inicialmente es importante destacar puntos de interés para un trabajo futuro:

- **Ampliación de datos.** El presente almacén cuenta con datos de alojamientos de Airbnb extraídos este año. No se dispone de un histórico de datos sobre ellos, como sí ocurre en el caso de la información proporcionada por el INE. Sería muy interesante ampliar la información de los alojamientos a lo largo del tiempo, para ver así una evolución de los mismos a través de los años: comprobar si aumentan o descienden sus precios, comparar las evaluaciones recibidas, ver si se añaden servicios, etc.

- **Aumento de fuentes.** Otro punto de interés sería la incorporación de más fuentes de datos y la integración de esta nueva información en el almacén para la realización de nuevos análisis que permitan proporcionar un mayor conocimiento sobre las características existentes o información adicional sobre características nuevas.

- **Predicciones.** También sería relevante incorporar nuevos modelos de predicciones para otros valores, como pueden ser predecir las mejores ubicaciones para pisos con determinadas características o estimar como evolucionará la economía turística del país a lo largo del tiempo, así como ir mejorando los modelos a partir de ajustes de parámetros e incorporación de datos futuros.

8.2 Valoración personal

Para finalizar, me gustaría destacar la oportunidad que me ha dado la realización de este proyecto de poner en práctica diferentes conocimientos y habilidades adquiridas a lo largo de la carrera, así como de constatar su valor y funcionalidad para mi futuro profesional.

Apéndices

Apéndice A

Script utilizado para la creación de tablas

```
CREATE TABLE `alojamientos` (
  `id_alojamiento` int NOT NULL,
  `name` varchar(250) DEFAULT NULL,
  `viajeros` double DEFAULT NULL,
  `media_ratings` double DEFAULT NULL,
  `n_evals` int DEFAULT NULL,
  `amenities` varchar(300) DEFAULT NULL,
  `lat` double DEFAULT NULL,
  `lng` double DEFAULT NULL,
  `Dormitorios` double DEFAULT NULL,
  `Camas` double DEFAULT NULL,
  `Baños` double DEFAULT NULL,
  `Idiomas` varchar(200) DEFAULT NULL,
  `Ratio de respuesta` varchar(4) DEFAULT NULL,
  `Tiempo de respuesta` varchar(20) DEFAULT NULL,
  `Hora_entrada` time DEFAULT NULL,
  `Hora_salida` time DEFAULT NULL,
  `id_mun` int DEFAULT NULL,
  `año_registro` int DEFAULT NULL,
  `mes_registro` int DEFAULT NULL,
  `num_amenities` int DEFAULT NULL,
  `num_idiomas` int DEFAULT NULL,
  `internet` varchar(3) DEFAULT NULL,
  `vistas_panoramicas` varchar(3) DEFAULT NULL,
  `calefaccion_refrigeracion` varchar(3) DEFAULT NULL,
  `aparcamiento` varchar(3) DEFAULT NULL,
  PRIMARY KEY (`id_alojamiento`)
)

CREATE TABLE `alquiler` (
```

```
  `id_alojamiento` int DEFAULT NULL,
  `id_anfitrion` int DEFAULT NULL,
  `id_cancelacion` int DEFAULT NULL,
  `año_extraccion` int DEFAULT NULL,
  `mes_extraccion` int DEFAULT NULL,
  `price_per_night` double DEFAULT NULL
)


CREATE TABLE `anfitriones` (
  `id_anfitrion` int NOT NULL,
  `superanfitrion` varchar(2) DEFAULT NULL,
  PRIMARY KEY (`id_anfitrion`)
)

CREATE TABLE `cancelacion` (
  `id_cancelacion` int NOT NULL,
  `tipo_cancelacion` varchar(25) DEFAULT NULL,
  PRIMARY KEY (`id_cancelacion`)
)

CREATE TABLE `ccaa` (
  `name` varchar(76) DEFAULT NULL,
  `id_ccaa` int NOT NULL,
  PRIMARY KEY (`id_ccaa`)
)


CREATE TABLE `municipios` (
  `id_mun` int NOT NULL,
  `id_ccaa` int DEFAULT NULL,
  `prov_code` int DEFAULT NULL,
  `prov_name` varchar(76) DEFAULT NULL,
  `mun_name` varchar(100) DEFAULT NULL,
  PRIMARY KEY (`id_mun`)
)

CREATE TABLE `fecha` (
  `id_fecha` int NOT NULL,
  `mes` int DEFAULT NULL,
  `año` int DEFAULT NULL,
  PRIMARY KEY (`id_fecha`)
)


CREATE TABLE `ine` (
```

```
  `id_ccaa` int DEFAULT NULL,
  `gasto_total` double DEFAULT NULL,
  `gasto_medio_persona` double DEFAULT NULL,
  `gasto_medio_diario_persona` double DEFAULT NULL,
  `duracion_media_viaje` double DEFAULT NULL,
  `n_turistas` double DEFAULT NULL,
  `periodo` int DEFAULT NULL
)

CREATE TABLE `gasto_total` (
  `VALOR` double DEFAULT NULL,
  `año` int DEFAULT NULL,
  `mes` int DEFAULT NULL
)
```

Apéndice B

Script utilizado para la creación de claves foráneas

```
-- Crear la restricción de clave foránea en la columna 'id_mun' de
    la tabla 'alojamientos'
ALTER TABLE alojamientos
ADD CONSTRAINT fk_id_mun
FOREIGN KEY (id_mun)
REFERENCES municipios(id_mun);

-- Crear la restricción de clave foránea en la columna
    'id_fecha_registro' de la tabla 'alojamientos'
ALTER TABLE alojamientos
ADD CONSTRAINT fk_id_fecha_registro
FOREIGN KEY (id_fecha_registro)
REFERENCES fecha(id_fecha);


-- Crear las restricciónes de clave foránea en la tabla 'alquiler'
ALTER TABLE alquiler
ADD CONSTRAINT fk_id_alojamiento
FOREIGN KEY (id_alojamiento)
REFERENCES alojamientos(id_alojamiento);

ALTER TABLE alquiler
ADD CONSTRAINT fk_id_fecha
FOREIGN KEY (id_fecha)
REFERENCES fecha(id_fecha);

ALTER TABLE alquiler
ADD CONSTRAINT fk_id_cancelacion
FOREIGN KEY (id_cancelacion)
REFERENCES cancelacion(id_cancelacion);
```

```
29
30 ALTER TABLE alquiler
31 ADD CONSTRAINT fk_id_anfitrion
32 FOREIGN KEY (id_anfitrion)
33 REFERENCES anfitriones(id_anfitrion);
34
35 -- Crear la restricción de clave foránea en la columna 'id_fecha'
       de la tabla 'gasto_total'
36 ALTER TABLE gasto_total
37 ADD CONSTRAINT fk_fecha_gasto
38 FOREIGN KEY (id_fecha_gasto)
39 REFERENCES fecha(id_fecha);
40
41
42 -- Crear la restricción de clave foránea en la columna 'id_ccaa' de
       la tabla 'municipios'
43 ALTER TABLE municipios
44 ADD CONSTRAINT fk_id_ccaa
45 FOREIGN KEY (id_ccaa)
46 REFERENCES ccaa(id_ccaa);
47
48 -- Crear la restricción de clave foránea en la columna
       'id_fecha_periodo' de la tabla 'ine'
49 ALTER TABLE ine
50 ADD CONSTRAINT fk_fecha_periodo
51 FOREIGN KEY (id_fecha_periodo)
52 REFERENCES fecha(id_fecha);
53
54 -- Crear la restricción de clave foránea en la columna 'id_ccaa' de
       la tabla 'ine'
55 ALTER TABLE ine
56 ADD CONSTRAINT fk_ccaa
57 FOREIGN KEY (id_ccaa)
58 REFERENCES ccaa(id_ccaa);
```

Lista de acrónimos

ETL Extraction, Transformation and Load. i, ii, 2, 3, 6, 7, 28–56, 68

INE Instituto Nacional de Estadística. ii, 1, 4, 18, 19, 22, 23, 25, 26, 28, 42, 53, 59, 60

Bibliografía

[1] "Airbnb: qué es y cómo funciona," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.airbnb.es/help/article/2503

[2] "Pon tu casa en Airbnb," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.airbnb.es/host/homes

[3] "Instituto Nacional de Estadística," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.ine.es/ss/SatelliteL=es_ES&c=Page&cid=1254735910183&p=1254735910183&pagename=INE%2FINELayout

[4] R. Kimball and J. Caserta, "The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data," 2004, consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.safaribooksonline.com/library/view/the-data-warehouse/9780764567575/

[5] E. Carisio, "ETL: todo sobre el proceso de Extract, Transform and Load," 2022, consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://blog.mdcloud.es/que-es-etl-extraccion-transformacion-y-carga/

[6] "¿Qué es un almacén de datos?" consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.oracle.com/es/database/what-is-a-data-warehouse/#link2

[7] R. D. Thantilage, N.-A. Le-Khac, and M.-T. Kechadi, "Healthcare data security and privacy in Data Warehouse architectures," 2023, consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S2352914823001144#sec3

[8] "Research on Big Data Analysis Data Acquisition and Data Analysis," 2021, consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://ieeexplore.ieee.org/document/9545987

[9] "Diferentes tipos de análisis estadístico de datos," 2022, consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.esic.edu/rethink/tecnologia/tipos-de-analisis-estadistico-de-datos

[10] "Conozca la familia de Visual Studio," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://visualstudio.microsoft.com/es/

[11] "What is MySQL?" consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://dev.mysql.com/doc/refman/8.0/en/what-is-mysql.html

[12] "Consiga que sus datos tengan un efecto inmediato," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://powerbi.microsoft.com/es-es/

[13] "Espacio de recursos de ciencia de datos, R Studio," consultado el11 de septiembre de 2023. [En línea]. Disponible en: http://datascience.recursos.uoc.edu/es/r-studio/

[14] "RSTUDIO IDE, The most trusted IDE for open source data science," consultado el11 de septiembre de 2023. [En línea]. Disponible en: https://posit.co/products/open-source/rstudio/

[15] "About Python," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.python.org/about/

[16] "¿Qué es Python?" consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/python/

[17] "Beautiful Soup Documentation," consultado el11 de septiembre de 2023. [En línea]. Disponible en: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[18] "Selenium with Python," consultado el11 de septiembre de 2023. [En línea]. Disponible en: https://selenium-python.readthedocs.io/

[19] "About Pandas," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://pandas.pydata.org/about/index.html

[20] "GeoJSON Municipios España, OpenDataSoft," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://data.opendatasoft.com/explore/dataset/georef-spain-municipio%40public/information/?disjunctive.acom_code&disjunctive.acom_name&disjunctive.prov_code&disjunctive.prov_name&disjunctive.mun_code&disjunctive.mun_name&dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6Imdlb3JlZi1zcGFpbi1tdW5pY2lwaW9AcHVibGl
location=7,40.43022,-5.34485&basemap=jawg.streets

[21] "Número de turistas según comunidad autónoma de destino principal, INE," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.ine.es/jaxiT3/Tabla.htm?t=23988

[22] "Gasto de los turistas según comunidad autónoma de destino principal, INE," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.ine.es/jaxiT3/Tabla.htm?t=23998&L=0

[23] "Gasto total mensual en turismo, INE," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://www.ine.es/consul/serie.do?d=true&s=FREG375

[24] I. Smirnov, "Educational project about Data Science," consultado el 11 de septiembre de 2023. [En línea]. Disponible en: https://github.com/x-technology/airbnb-analytics