Practica 1

Análisis de precios de vuelos nacionales

Contexto

Una de las actividades más comunes hoy en día es realizar viajes, ya sea con amigos o en familia. Para llegar al destino del viaje en su mayoría se utiliza como método de transporte el avión y, como en todo, queremos el mejor precio. A pesar de que hay comparadores de vuelos por todo lo ancho de la web, nos encontramos que no sabemos realmente cuál es el mejor día para comprar un vuelo. Por ello, el objetivo de esta práctica es comprobar si el precio de los vuelos varía según el día de la semana en que se realiza la búsqueda.

Hemos acotado la búsqueda a vuelos nacionales que parten del aeropuerto de Madrid-Barajas. Por ello, hemos recogido vía web scraping las compañías que operan en ese aeropuerto a través de la web de AENA, y las hemos comparado con los aeropuertos que hay en España que hemos extraído de Wikipedia. Estas dos extracciones de datos se pueden considerar estáticas.

Hemos escogido el comparador de vuelos de Google para realizar las consultas dinámicas con los diferentes destinos que hemos recabado de las otras páginas web. Este comparador se llama Google Flights y os dejamos el enlace de la comparación de vuelos entre Madrid y Barcelona: click aquí para ir al enlace.

Descripción del dataset

El dataset que se presenta en esta práctica extrae todos aquellos datos que el usuario ve relevante a la hora de comprar un billete de avión, como puede ser el precio o las escalas. Los datos los hemos acotado para no generar una base de datos enorme, por ello hemos escogido un único aeropuerto, que en este caso es el de Madrid-Barajas. Y hemos recogido datos de todos los vuelos que parten hacia destinos nacionales.

El resultado ha sido una base de datos bastante completa, a esperas de ser filtrada y procesada, pero con potencial para ser analizada.

Contenido

Los datos que se presentan en este dataset representan datos desde el domingo 23/04/2023 al martes 25/04/2023.

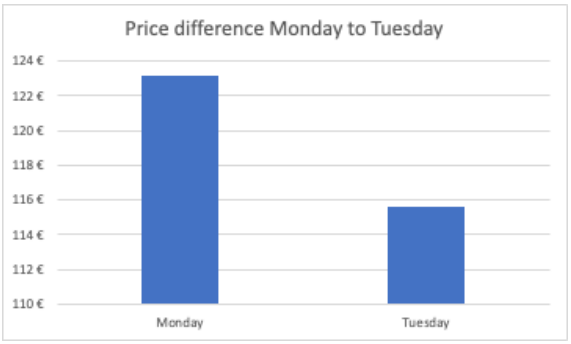
Estos datos se han extraído del apartado de vuelos de google, que ya se encarga de comparar las diferentes compañías y mostrarlas en formato web.

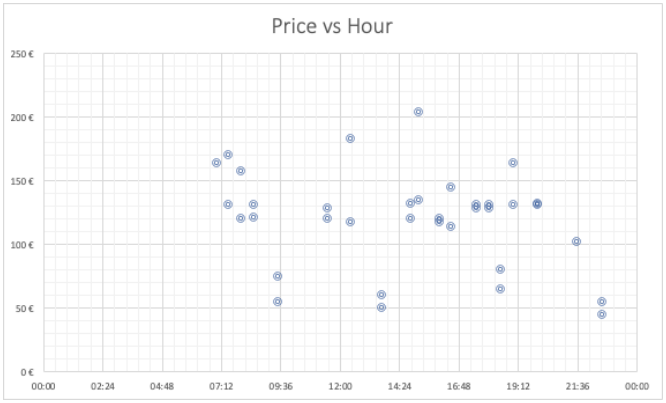
El dataset que hemos construido consta de 10 columnas.

La **primera columna** trata de la aerolínea que opera el vuelo.
La **segunda columna** da el precio que cuesta el vuelo al destino que se encuentra en la décima columna.
En la **tercera columna** se encuentra la duración del viaje.
La **cuarta columna** indica si el vuelo es directo o si tiene alguna escala.
En la **quinta y sexta** columna se indica la **hora de salida** del avión y de **llegada**.
En la **séptima columna** se indica si el vuelo es de ida o de ida y vuelta.
En la **octava columna** se indica el tiempo en el que se recogió el dato.
En la **novena columna** se encuentra el origen del vuelo, que siempre será Madrid, aquellos datos que se salgan de este campo quedarán descartados.

Airline	Price	Duration	Stops	Departure	Arrival	Returns	Datetime	From	To

Representación gráfica





Propietario

El propietario de estos datos es de las compañías de vuelos que operan en el aeropuerto de Madrid, como por ejemplo, Iberia, Air Europa o Ryanair. El propietario de juntar toda esta información es Google, el cuál la presenta.

Debido a que estos datos son de dominio público y que no representan la adquisición de ningún dato sensible, ni personalizado de usuario, no supone ninguna contradicción ética. Estos datos permiten analizar al usuario de una forma mucho más precisa cómo varía el precio de los vuelos a lo largo del tiempo.

Inspiración

El conjunto de datos es interesante para todos los usuarios que viven en Madrid y alrededores, para ver ofertas que le pueden salir para una escapada a algún rincón de España. Además le hará conocedor de que días es el mejor para realizar una compra de billetes.

La pregunta que tratamos de responder es si el precio del mismo billete cambia según el día de la semana en que se compre.

Licencia

Debido a que los datos son de dominio público y todos pueden acceder a ellos sin restricciones, se considera que la base de datos que adjuntamos no debe tener ninguna licencia y debe ser de dominio público para que cualquier persona se pueda beneficiar.

Released Under CC0: Public Domain License.

Codigo

Respecto al código las principales complicaciones con las que nos hemos encontrado son que las url de los principales comparadores de vuelos o viajes se generan de forma diferente en cada búsqueda por lo que dificulta el realizar la búsqueda de forma automática.

Además que en algunas páginas si es la primera vez que se accede a ellas piden aceptar las cookies cosa que impide, en alguna ocasiones, recuperar datos de la página mediante código.

La forma en la que hemos solventado estos problemas es utilizando Selenium y las búsquedas de Google, junto con el comparador de vuelos de google.

La ventaja de utilizar Google es que puedes repetir la búsqueda con los mismos criterios de forma sencilla y automática. Además que al utilizar el comparador de vuelos de Google ya tenemos la información de varios de los comparadores más populares de una sola búsqueda.

Por otro lado el utilizar Selenium nos permite mucha más flexibilidad a la hora de automatizar la búsqueda y navegación por diferentes páginas web.

Selenium es una herramienta open source que permite realizar test automáticos en páginas web. Nosotros lo hemos utilizado para poder aceptar las cookies de Google para luego realizar la búsqueda y acceder a la página del comparador de vuelos de Google. Una vez que ya hemos conseguido cargar la lista de todos los vuelos los recuperamos y guardamos en un archivo .csv.

También es importante destacar que no todas las búsquedas se han hecho de esta forma ya que recuperamos datos de más sitios web que se pueden realizar de forma estática.

Por lo que la secuencia en la que nuestro programa recupera los datos es la siguiente:

1. Recupera los datos de los destinos a los que se puede viajar desde Madrid en la página de Aena. Esta búsqueda es estática y se hace mediante la librería `requests` de Python.
2. En la búsqueda anterior solo podemos obtener los códigos IATA de los destinos. El código IATA es un código de tres letras que hace referencia a la ciudad o región donde opera un aeropuerto. Con este dato no podíamos obtener los resultados que deseábamos en las búsquedas por lo que necesitábamos el nombre de la ciudad o aeropuerto. Por lo que lo siguiente que hacemos es recuperar desde wikipedia la relación del código IATA con la ciudad. Esta búsqueda también es estática y la seguimos realizando con la librería `request`.
3. Por último, ya podemos hacer la búsqueda del vuelo siendo el origen Madrid y el destino las ciudades con las que tenemos conexión desde el aeropuerto de Barajas.

El utilizar Selenium nos ha obligado a tener en consideración ciertos aspectos:

- Selenium abre el navegador web, en este caso en concreto hemos utilizado Chrome, por lo que sí tenemos la sesión iniciada y es un navegador que utilicemos habitualmente, lo normal es que ya hayamos aceptado las políticas de Google. Esto puede provocar que en algunos ordenadores funcione bien el programa y en otros no. Para solucionar esto hemos decido abrir el navegador en modo incógnito. Así da igual en qué ordenador se ejecute el programa que siempre pedirá primero aceptar las cookies.
- Otro problema con el que nos hemos encontrado es que para poder aceptar las cookies es necesario encontrar el botón y para ello tenemos que buscar el conjunto de palabras: “Aceptar todo” pero esto solo se cumple si el navegador lo tenemos configurado en español. Para solucionar esto una de las primeras comprobaciones que se hace es el idioma en el que está el navegador.
- Por último, antes de poder realizar la extracción de los datos hemos tenido que añadir un delay de unos 10 segundos ya que en algunas de las búsquedas en las que se obtienen muchos resultados tardaba un poco en cargarlas. De esta forma evitamos que nos saltará una excepción al recuperar la lista completa de vuelos.

Dataset

Zenodo: https://zenodo.org/record/7864245#.ZEgOv3ZBzao

Vídeo

Vídeo en google drive:
https://drive.google.com/file/d/1oQI8tovfvXIFByTRAKNRyy7o6Qm4VKGW/view?usp=sharing

Contribuciones

Contribuciones	**Firma**
Investigación previa	DC, PM
Redacción de las respuestas	DC, PM
Desarrollo del código	DC, PM
Participación en el vídeo	DC, PM