

EDA - Invertir en un piso de alquiler vacacional en Málaga

Andrada Marcusanu

Bootcamp Data Science nov.2023 – The Bridge

El tema del proyecto de análisis exploratorio de datos surge del conocimiento personal que tengo sobre la ciudad de Málaga y su mercado de alquiler turístico. Corroborado con datos del INE sobre la distribución de viviendas vacacionales, resultó entonces un camino interesante para explorar.

Considero que lo más importante en un análisis es la calidad de los datos. Estos deberían ser, de forma ideal, completos, suficientes y actualizados. Por esta razón he optado por la estrategia del web scraping, así pudiendo obtener resultados de última hora y, quizás, realizar también un seguimiento de ellos a lo largo del tiempo.

1. OBTENCION DE DATOS

1.1. Webscraping de AirDNA

AirDNA es una plataforma que proporciona estadísticas sobre los alquileres turístico del mundo entero y que, a su vez, se basa en el web scraping de Airbnb y VRBO. Lleva al menos 5 años en el mercado y en todo este tiempo estuvo registrando los datos de estas dos grandes plataformas de alquiler turístico para poder realizar estadísticas y arrojar luz sobre la rentabilidad y alcance de las mismas.

Scrapeando se puede sacar información como:

- Nombre del anuncio y número de referencia de Airbnb/VRBO
- Número de habitaciones, baños, y huéspedes
- Rating y número de reviews
- Precio medio diario de alquiler
- Ganancia media anual (últimos 12 meses)
- Ocupación media (últimos 12 meses)
- Ganancia potencial esperada
- Ubicación (latitud y longitud)
- Imágenes, etc.

La estrategia que utilicé ha sido el **webscraping con Selenium** y el uso de la **librería keyboard de Python** para poder navegar el panel de Chrome Dev Tools y descargar directamente de allí un set de respuestas .JSON que contienen los datos en el mejor formato posible y proporcionan alguna información que se debería obtener solo a través de una cuenta de pago.

Inicialmente había realizado un web scraping con Selenium sin abrir el panel DevTools, buscando únicamente en la web, que se genera de forma dinámica. Este proceso era muy largo (duraba hasta **8 horas** para mostrar y scrapear los aprox. 10.300 anuncios de Málaga).

Para recortarlo apliqué un filtro: sabiendo que la web me muestra los anuncios en orden descendiente, del anuncio con más reviews al anuncio con menos, y estimando que un anuncio que lleva mínimo 12 meses en el mercado (ya que los otros no serían relevantes para el estudio) debe tener mínimo 50 reviews, he introducido una clausula de que el bot se pare cuando encuentra un anuncio con 49 reviews. Esto ha reducido el tiempo de scrapeo a **1.5-2 horas** para aprox. 2300 anuncios.

No obstante, al descubrir el recurso de respuestas JSON, he preferido realizarlo de esta manera ya que reducía el tiempo de scrapear los 2300 anuncios a **27 minutos**.

Un desafío ha sido asignarle un distrito de Málaga a cada anuncio ya que solo tenía como dato latitud y longitud.

Para ello he utilizado la **red de datos abierta del Ayuntamiento de Málaga** que proporciona archivos geoJSON con los polígonos que definen el perímetro de los 11 distritos de la ciudad.

1.2. Webscraping de idealista

Elegí idealista por ser la mayor plataforma de venta y alquiler de toda España. Realicé un **webscraping rápido**, sacando manualmente los 11 url-s de los 11 distritos de Málaga (para tener ubicados los anuncios dentro de cada distrito) y luego utilizando **BeautifulSoup** para iterar por todas las páginas siguientes de cada una de las 11 url-s, hasta sacar todos los anuncios de la ciudad de Málaga, por sus 11 distritos, y finalmente juntar todos los datos en un .csv

2. ANALISIS EXPLORATORIO DE DATOS

El EDA comienza con **la limpieza de los datos** (formato, eliminar caracteres innecesarios, valores nulos, pasar de string a int, eliminar columnas innecesarias, etc.), que se realiza en un notebook aparte.

En otro notebook, llamado main.ipynb, se realizaron los **gráficos de análisis**, la mayoría tipo univariante (bar chart, pie chart, density plot, histogram), algunos multivariante (diagrama de dispersión, matrices de correlación) y algunos mapas.

3. CONCLUSIONES

Las hipótesis iniciales fueron:

- Estimamos que invertir en una ciudad como Málaga podría ser una inversión con una buena rentabilidad
- También estimamos que se podría realizar una inversión rentable con un presupuesto no muy elevado, de máximo 200.000 Euros
- Y, por último, estimamos que la zona en la que conviene adquirir la vivienda sería el Centro

Y las conclusiones:

- La TIR media de la tabla adyacente se sitúa en torno al 5.75%, lo que se podría traducir en una mejor o peor rentabilidad según los parámetros de la inversión inicial.
- No obstante, se pueden obtener rentabilidades TIR cercanas o superiores al 10%
- Cabe destacar que dos de las inversiones que proporcionan mayor rentabilidad (invertir en un estudio en Cruz de Humilladero o en Bailén-Miraflores) cuentan también con una inversión inicial por debajo de los 200.000 Euros
- A pesar de que el distrito con más oferta sea el Centro y que el más caro sea Este, las mejores inversiones parecen estar en los distritos adyacentes y en viviendas de menor tamaño

Las primeras dos hipótesis se han confirmado, pero en cambio la tercera no, los datos sugiriendo que habría que mirar de cerca barrios adyacentes al Centro que podrían tener mucha mayor rentabilidad.

También descubrimos que **las viviendas de menor tamaño suelen ser más rentables que las viviendas grandes.**