Proyecto hosting

El proyecto hosting tiene como objetivo principal terminar de unir los conceptos para el alumno de forma práctica mediante un proyecto.

Introducción

El proyecto es para WhiteHosting, una empresa británica que se dedica a la inversión en inmuebles vacacionales. BlackWidow está etiquetada como uno de los mayores fondos de occidente y ha visto la etapa post covid como una potencial oportunidad. WhiteHosting tiene inmuebles en España y por tanto conoce muy bien el mercado español.

WhiteHosting nos ofrece los siguientes datos:

- Dataset en csv scrapeado
- Documentación para hacer un bot de scrapeo a airbnb

¿Qué aprenderás en el proyecto?

- Web Scraping
    - Beautiful soap
    - Selenium
- Limpieza de datos y análisis preliminar
    - Pandas
        - Numpy
        - Funciones avanzadas de python
- Creación de hipótesis
- Presentación de los resultados y sugerencias en herramienta de BI
    - Tableau
    - Power BI

Herramientas a usar en el proyecto

- **Python:** es nuestro principal lenguaje para analizar datos y utilizar sus librerías nativas
- **Pandas:** utilizaremos pandas para agrupar, modelar y entender los datos
- **Conda :** como gestor de entornos
- **Power BI / Tableau:** Por último representaremos la información en cualquiera de ellas

Recursos complementarios para el proyecto

- Dataset mercado español
- Recursos para scraping de airbnb

Problema a resolver

WhiteHosting es consciente de que la era post pandemia puede ser una oportunidad de compra.

Al ser WhiteHosting experto en el mercado inmobiliario español , quiere invertir 300 millones de euros en alojamientos de Madrid , Barcelona y Valencia (solo capitales de provincia).

WhiteHosting espera obtener conclusiones de nuestro análisis para comprender mejor el mercado y las distintas oportunidades de compra.

El cliente espera un archivo ejecutable. Este archivo deberá ser un Jupyter Notebook o Google Colab que se usará para comprobar los resultados. En él deberá aparecer las respuestas a las preguntas de negocio, el código SQL, un esquema relaciones, todo el código de Python con su salida.

El team lead del proyecto nos indica que las fuentes de datos tienen la siguiente estructura:

Dataset escrapeado mercado español

Explicación tabla

Col_name	Example
apartment_id	36187629
md5	66fff4225feb2ddf104ea38f76e4bff1
name	Piso reformado excelente ubicacion
description	tamento reformado hace dos meses en el C

host_id	261787331
neighborhood_overview	NaN
neighbourhood_name	BETERO
neighbourhood_district	POBLATS MARITIMS
latitude	3.947.149
longitude	-3.346
room_type	Entire home/apt
accommodates	4
bathrooms	2.0
bedrooms	3.0
beds	3.0
amenities_list	V,Wifi,"Air conditioning",Kitchen,"Smoking a
price	90.0
minimum_nights	5
maximum_nights	20
has_availability	TRUE
availability_30	0
availability_60	4
availability_90	34
availability_365	34
number_of_reviews	1
first_review_date	2019-08-19
last_review_date	2019-08-19
review_scores_rating	100.0
review_scores_accuracy	10.0
review_scores_cleanliness	10.0
review_scores_checkin	10.0
review_scores_communication	10.0
review_scores_location	10.0
review_scores_value	10.0
license	NaN
is_instant_bookable	TRUE
reviews_per_month	0.22
country	spain
city	valencia
insert_date	2019-12-31

Instrucciones del scraping de Airbnb

Pasos

1. Scrapear paginas with Python and Beautiful Soup
2. Enfrentarse a páginas dinámicas
3. Paralelizar con multiprocesamiento

Documentacion : https://github.com/x-technology/airbnb-analytics

Case study Questions

BlackWidow quiere tener respuesta al menos a estas preguntas :

1. ¿Cuántos inmuebles únicos hay en el dataset vs airbnb?
2. ¿Cual es la diferencia en número de inmueble posteados entre los 2 datasets?
3. ¿Cuáles son los parámetros que forman el precio?
4. ¿Son iguales los parámetros en ambas fuentes de datos?
5. ¿Has podido ver algún inmueble que está en ambas fuentes de datos?, es su valoración la misma?

6. ¿En qué fuentes de datos hay más volatilidad en cuanto al precio? , hay algún tipo de estacionalidad?
7. **Crea un dataset único de verdad**
8. ¿Hay alguna variable que podríamos añadir para mejorar el análisis?
9. ¿Puedes hacer un ranking de los inmuebles más caros?
10. ¿Cuáles son los inmuebles más rentables?
11. Teniendo en cuenta todo el conocimiento adquirido , cuales son tus sugerencias a la hora de invertir 300 millones de euros.

Informe del proyecto

Deberás entregar un **resumen del proyecto** (en él no se debe plasmar todos los detalles ni el código ejecutado) a realizar en diapositivas que expondrán brevemente la **metodología** seguida en este informe. No debe superar las 5 páginas y debe tener los siguientes apartados:

o **Equipo del proyecto y objetivos**: Exponer brevemente los integrantes del equipo, las tareas desempeñadas por cada uno y los objetivos planteados para el trabajo.

o **Modelo relacional:** En él se debe exponer el modelo relacional diseñado e implementado (puede hacerse uso del esquema solicitado).

o **Limpieza de datos**: Breve exposición de los errores encontrados y los pasos a seguir en la limpieza de los datos.

Aparte de la presentación con la **metodología** del proyecto, deberás entregar lo siguiente:

o **Notebooks de Python**. Será necesario entregar los **notebooks** en los que se realiza la limpieza y tratamiento de datos. Estos servirán de apoyo en la presentación si se requieren detalles más técnicos sobre estos procesos. Es decir, no se enseñarán durante la presentación pero pueden aparecer en la ronda de preguntas. Se valorará la limpieza y la estructura lógica de los notebooks.

o **Respuestas** a las preguntas solicitadas. Para mostrar los resultados se usará una plataforma de Business Intelligence como **Tableau o Power BI**. Este informe será presentado a continuación de la metodología. De nuevo, el formato y la presentación serán muy importantes. Se valorará muy positivamente que los estudiantes sean capaces de formular y dar respuesta a nuevas preguntas que ayuden a entender mejor la temática de negocio planteada.