Desafío 1. Análisis exploratorio de un dataset de precios de propiedades

Introducción

La inmobiliaria Properati publica periódicamente información sobre ofertas de propiedades para venta y alquiler. Ud. deberá asesorar a la inmobiliaria a desarrollar un modelo de regresión que permita predecir **el precio por metro cuadrado de una propiedad**. El objetivo final, a concretarse en el Desafío 2, es que el modelo que desarrollen sea utilizado como tasador automático a ser aplicados a las próximas propiedades que sean comercializadas por la empresa. Para ello la empresa le provee de un dataset correspondiente al primer semestre de 2017.

El dataset es de tamaño entre pequeño y mediano, pero tiene dos complejidades a las que deberá prestarle atención:

- Peso de missing data en algunas variables relevantes.
- Será importante tener en cuenta el problema de la influencia espacial en los precios por metro cuadrado. En efecto, es probable que existan diferencias importantes en las diferentes geografías, barrios y zonas analizadas.

Objetivos del Desafío 1:

- Efectuar una limpieza del dataset provisto. Particularmente, deberá diseñar estrategias para lidiar con los datos perdidos en ciertas variables.
- Realizar un análisis descriptivo de las principales variables.
- Crear nuevas columnas a partir de las características dadas que puedan tener valor predictivo.

Datos

Descargar el dataset de:
https://drive.google.com/file/d/0BzVrTKc02N8qNUdDSExBQlFTNlU/view

Requisitos y material a entregar

1. Una jupyter notebook/lab que satisfaga los requerimientos del proyecto, donde se realice y muestre la limpieza y el análisis de los datos a ser entregados. La notebook deberá estar debidamente comentada, y debe poder ejecutarse sin errores de principio a fin.
2. Una exposición de no más de 10 minutos del trabajo realizado, consistente en una presentación acompañada con algunos slides no técnicos (PPT o Google Slides).

   La presentación debe constar de:

   - Una introducción (planteo del problema, la pregunta, la descripción del dataset, etc.)
   - Un desarrollo de los análisis realizados (análisis descriptivo, análisis de correlaciones preliminares, visualizaciones preliminares)
   - Una exposición de los principales resultados y conclusiones.

   Esta presentación se entregará como un video grabado, junto con el ppt/Google slides para que pueda ser revisado por la dupla docente.

Fecha de entrega

● El material deberá entregarse **como máximo el Martes 6 de abril**. El material incluye la/as jupyter notebooks/labs que hayan generado para hacer la limpieza, y el video en el que realizan la presentación del trabajo. **El día 12 de abril** se realizará una clase de devolución general de las presentaciones y las notebooks. .

Dataset

El dataset contiene información sobre todas las propiedades georeferenciadas de la base de datos de la empresa. La información de cada propiedad que incluye es la siguiente:

- ID de registro
- Tipo de la propiedad (house, apartment, ph)
- Operación del aviso (sell, rent)
- Nombre del lugar
- Nombre del lugar + nombre de sus ‘padres’
- ID de geonames del lugar (si está disponible)
- Latitud, longitud

- Precio original del aviso
- Moneda original del aviso (ARS, USD)
- Precio del aviso en moneda local (ARS)
- Precio aproximado en USD
- Superficie en m2
- Superficie cubierta en m2
- Precio en USD/m2
- Precio por m2
- N° de piso, si corresponde
- Ambientes
- URL en Properati
- Descripción
- Título
- URL de un thumbnail de la primera foto

¿Cómo empezar? Sugerencias

Agreguen toda otra información construida a partir de los datos originales (o incluso información externa) que consideren relevante y útil para resolver los objetivos planteados.

Aprovechen las herramientas de Pandas: *groupby*, *summation, pivot_tables* y otras aplicaciones y métodos de los DataFrames que hacen mucho más simples los cálculos y otras agregaciones de los datos.

En la presentación de los resultados tengan en cuenta que es altamente probable que la audiencia no tenga un nivel técnico, así que mantengan el lenguaje en un nivel accesible.

En términos generales, recuerden las siguientes sugerencias:

- Escribir un pseudocódigo antes de empezar a codear. Suele ser muy útil para darle un esquema y una lógica generales al análisis.
- Leer la documentación de cualquier tecnología o herramienta de análisis que usen. A veces no hay tutoriales para todo y los documentos y las ayudas son fundamentales para entender el funcionamiento de las herramientas utilizadas.
- Documentar todos los pasos, transformaciones, comandos y análisis que realicen.

Recursos útiles

- Documentación de la librería GeoPandas
- Cheatsheet de pandas
- Pandas user guide