Trabajo Práctico 1 : Propiedades en Venta

Introducción

En este trabajo práctico se propone que cada alumno se enfrente a un problema real de ciencia de datos, que trabaje en cada una de las etapas del proceso y que pueda resolverlo aplicando los contenidos que vemos en la materia.

Vamos a utilizar el conjunto de datos provisto por la empresa Properati correspondiente a anuncios de propiedades en venta de la República Argentina publicados durante el año 2021 . La información fue extraída desde BigQuery (producto de Google Cloud para consultar grandes volúmenes de datos) donde la empresa disponibiliza sus datasets con avisos de propiedades y desarrollos inmobiliarios que están y estuvieron publicados en Properati en todo Latinoamérica desde 2015 hasta dos meses atrás. Los datos se actualizan diariamente para mayor información pueden consultar el siguiente link.

El objetivo principal del trabajo será aplicar técnicas de análisis exploratorio y preprocesamiento de datos. En la sección enunciado se detallan los objetivos particulares.

Modalidad de entrega

Notebook

El trabajo debe ser realizado en una notebook de python, se espera que la misma contenga **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. La notebook debe respetar la siguiente nomenclatura :

TP1_NOMBRE_ALUMNO_ENTREGA

En el caso que sea estrictamente necesario entregar más de una notebook las mismas deben contar con una numeración correlativa manteniendo un orden lógico entre ellas (TP1_NOMBRE_ALUMNO_ENTREGA_**N1**, TP1_NOMBRE_ALUMNO_ENTREGA_**N2**, etc) Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de markdown. Se debe incluir una sección principal con el título del trabajo y el nombre del alumno.

Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega.

Visualizaciones

Todos los gráficos que se incorporen deben tener su correspondiente título, leyenda, nombres en los ejes, unidades de medidas, y cualquier referencia que se considere necesaria. Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema, por lo tanto, deben ser comprensibles por quien los vaya a leer.

Preprocesamiento:
A partir de las tareas de preprocesamiento, y de las diferentes estrategias que se planteen, es posible que se generen nuevos datasets.

Repositorio
Cada alumno puede crear su propio repositorio en github con la siguiente nomenclatura:

CUDI-TP1-NOMBRE-ALUMNO

En dicho repositorio deberá estar disponible la notebook y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

Fechas de entrega: 2 semanas

Enunciado

El conjunto de datos a utilizar **properati_argentina_2021** se encuentra disponible en el siguiente enlace, la descripción de las variables se encuentra disponible aquí .

A continuación se detallan las etapas que deben ser desarrolladas en el trabajo: **Análisis Exploratorio y Preprocesamiento de Datos**

El primer paso consiste en la selección de los datos que se van a utilizar, se deben filtrar únicamente los anuncios de propiedades de tipo vivienda (Casa, PH y Departamento) ubicados en Capital Federal cuyo tipo de operación sea venta y su precio se encuentre en dólares (USD).

a) **Exploración Inicial :** analizar cada variable, considerando los siguientes aspectos
   - Tipo de variable
   - Variables Cuantitativas: calcular medidas de resumen: media, mediana, q1, q3,moda
   - Variables Cualitativas mostrar cantidad de valores posibles, y frecuencias de cada uno.
   - Determinar variables irrelevantes en el análisis (Ids por ejemplo)
   - Realizar un análisis gráfico de las distribuciones de las variables
   - Analizar las correlaciones existentes entre las variables.

   A partir de este análisis generar conclusiones sobre los datos.

b) **Datos Faltantes :** analizar la presencia de datos faltantes en el dataset
   - Realizar análisis de datos faltantes a nivel de columna. Graficar para cada variable el porcentaje de datos faltantes con respecto al total del dataset
   - Realizar un análisis de datos faltantes a nivel de fila. Calcular el porcentaje de datos faltantes de cada registro. Realizar un gráfico que permita conocer la proporción de faltantes por fila en el dataset.
   - Determinar, de ser posible, estrategias para reparar los valores faltantes.

- En caso de realizar imputaciones comparar las distribuciones de cada atributo reparado con la distribución anterior a la imputación de los datos faltantes.

c) **Valores atípicos :** analizar la existencia de valores atípicos
   - Detectar valores atípicos. Realizar gráficos que permitan visualizar los valores atípicos.
   - Explicar qué características poseen los datos atípicos detectados.
   - Decidir el tratamiento a aplicar sobre los mismos.
   - Analizar la relación entre el precio de venta y los metros de superficie ¿hay valores atípicos que no se detectaron previamente?

Conclusiones

Realizar las conclusiones correspondientes al trabajo realizado en su totalidad, destacando principalmente los aspectos que consideren más relevantes. Comentar brevemente qué otras opciones hubiesen explorado y quedaron fuera del alcance de este trabajo.