Datos duplicados

Waldir Toscano, Mausel Perez, Jorge Acosta

06/03/2024

Resumen

Este informe presenta un análisis exhaustivo de datos inmobiliarios de Medellín, Colombia, con un enfoque específico en la identificación y procesamiento de registros duplicados. Se detalla el proceso de limpieza de datos, transformación y agrupación realizado mediante un script en Python utilizando la biblioteca Pandas. Los resultados muestran la frecuencia de ocurrencia de cada registro único, lo que permite identificar patrones de repetición en los datos y facilita su posterior procesamiento. Este análisis es fundamental para garantizar la integridad de los datos en análisis posteriores del mercado inmobiliario.

1. Introducción

El análisis de datos inmobiliarios requiere una base de información limpia y confiable para garantizar la validez de las conclusiones derivadas. Uno de los problemas más comunes en las bases de datos inmobiliarias es la presencia de registros duplicados, que pueden distorsionar significativamente los resultados de cualquier análisis estadístico o modelo predictivo.

Este informe documenta el proceso de análisis de un conjunto de datos de propiedades inmobiliarias en Medellín, con énfasis en la detección de registros duplicados. El dataset original contiene información sobre diversos atributos de propiedades, como número de habitaciones, baños, estrato socioeconómico, y precios, entre otros.

La presencia de duplicados en estos datos puede deberse a múltiples factores:

- Propiedades idénticas listadas múltiples veces
- Errores en la captura o ingreso de datos
- Sincronización incorrecta entre diferentes sistemas de gestión inmobiliaria
- Actualización de registros sin eliminar versiones anteriores

El trabajo presentado aquí establece una metodología sistemática para identificar estos duplicados, cuantificarlos y prepararlos para su posterior procesamiento, lo que constituye un paso crítico en la preparación de datos para análisis inmobiliarios más sofisticados.

2. Objetivos del Análisis

El análisis tiene los siguientes objetivos principales:

1. **Preparar y limpiar el conjunto de datos**: Transformar los datos brutos en un formato adecuado para el análisis mediante la eliminación de caracteres no deseados y la estandarización de formatos numéricos.
2. **Identificar valores faltantes o problemáticos**: Cuantificar los valores que no pueden ser convertidos correctamente a formatos numéricos y desarrollar una estrategia para su manejo.
3. **Detectar registros duplicados**: Agrupar registros idénticos en todas sus características y cuantificar su frecuencia de aparición en el conjunto de datos.
4. **Crear un conjunto de datos procesado**: Generar un nuevo DataFrame que contenga los registros únicos junto con su frecuencia de ocurrencia para facilitar análisis posteriores.
5. **Evaluar la calidad de los datos**: Proporcionar métricas sobre la redundancia en el conjunto de datos para informar decisiones futuras sobre la recolección y gestión de datos inmobiliarios.

3. Descripción de los Análisis

El proceso de análisis implementado se divide en varias etapas clave, todas ejecutadas mediante un script de Python utilizando la biblioteca Pandas. A continuación se describe cada fase del proceso:

3.1. Carga y Exploración Inicial de Datos

Los datos se cargan desde un archivo CSV ubicado en la ruta "../../../Datasets/houses_medellin.csv". Durante la carga, se excluye la primera columna, que presumiblemente contiene identificadores o índices que no son relevantes para el análisis de duplicados.

```
datos = pd.read_csv(ruta_archivo_csv, index_col=0)
```

Una exploración inicial muestra las primeras filas del DataFrame para entender la estructura general de los datos antes de realizar cualquier modificación.

3.2. Renombrado de Columnas

Para facilitar el manejo de los datos, se renombran las columnas a un formato estandarizado, eliminando caracteres especiales y asignando nombres descriptivos en español:

```
datos.columns = [
    "Habitaciones", "Ba os", "Estrato", "Antig edad", "Piso_N", "
    Administracion",
    "Precio_m2", "Parqueaderos", "Estado", "Tipo_apartamento", "
    Precio",
    "Area_construida_m2", "Area_privada_m2"
]
```

Esta etapa garantiza la consistencia en la nomenclatura y facilita las referencias a columnas específicas en los análisis posteriores.

3.3. Limpieza y Transformación de Datos Numéricos

Una parte crítica del análisis es la normalización de valores numéricos. El script identifica varias columnas que deberían contener valores numéricos y aplica las siguientes transformaciones:

```
columnas_numericas = ["Habitaciones", "Ba os", "Estrato", "Precio_m2
    ",
    "Parqueaderos", "Precio", "Area_construida_m2", "Area_privada_m2"
    ]

for columna in columnas_numericas:
    # Eliminar caracteres no num ricos
    datos[columna] = datos[columna].astype(str).str.replace(r"
    [^0-9,.-]", "", regex=True)
    # Convertir comas a puntos decimales
    datos[columna] = datos[columna].str.replace(",", ".", regex=True)
    # Convertir a formato num rico
    datos[columna] = pd.to_numeric(datos[columna], errors="coerce")
```

Este proceso realiza tres acciones importantes:

- Elimina caracteres no numéricos (como símbolos de moneda o unidades)
- Convierte comas decimales a puntos decimales para estandarización
- Transforma las cadenas resultantes a valores numéricos, generando NaN cuando la conversión no es posible

3.4. Manejo de Valores Faltantes

El análisis cuantifica los valores que no pudieron ser convertidos a formato numérico (NaN) y adopta una estrategia para su manejo:

```
# Cuantificar valores NaN
print("\  n     Cantidad de NaN despu s de conversi n  num rica:")
print(datos[columnas_numericas].isna().sum())

# Reemplazar NaN con un valor indicador (-1)
datos[columnas_numericas] = datos[columnas_numericas].fillna(-1)
```

En lugar de eliminar registros con valores faltantes, se reemplazan con -1, un valor indicador que permite mantener estos registros en el análisis mientras se distinguen claramente de los valores reales.

3.5. Agrupación y Cuantificación de Duplicados

Finalmente, se realiza la agrupación de registros idénticos y se cuantifica su frecuencia:

```
# Agrupar por todas las columnas y contar frecuencia
datos_agrupados = datos.groupby(list(datos.columns)).size().
    reset_index(name="Frecuencia")
```

Esta operación es el núcleo del análisis de duplicados:

- Considera todas las columnas para la agrupación, identificando registros que son idénticos en todos sus atributos
- Cuenta la frecuencia de cada combinación única de valores
- Genera un nuevo DataFrame que incluye cada registro único junto con su frecuencia

4. Interpretación de Resultados

Los resultados del análisis proporcionan información valiosa sobre la calidad y estructura del conjunto de datos inmobiliarios:

4.1. Estructura y Limpieza de Datos

La fase de limpieza y transformación revela varios aspectos importantes:

- **Inconsistencia en formatos numéricos**: La necesidad de eliminar caracteres no numéricos y convertir separadores decimales indica que los datos originales presentaban inconsistencias en la representación de valores numéricos.
- **Presencia de valores no convertibles**: La cuantificación de valores NaN después de la conversión numérica sugiere que algunas entradas contenían datos que no podían interpretarse como números, lo que podría indicar errores de entrada o categorías codificadas incorrectamente como numéricas.
- **Estrategia de conservación de datos**: Al reemplazar valores NaN con -1 en lugar de eliminar registros, el análisis prioriza la conservación de datos, permitiendo un examen más completo de posibles duplicados incluso en presencia de valores problemáticos.

4.2. Análisis de Duplicados

El resultado principal del análisis es el DataFrame agrupado que contiene:

- Cada combinación única de valores para todas las columnas
- La frecuencia con que cada combinación aparece en el conjunto original

De este resultado podemos interpretar:

- **Distribución de duplicados**: Registros con frecuencia mayor a 1 representan duplicados en el conjunto de datos. La distribución de estas frecuencias indica el grado de redundancia en los datos.
- **Patrones de duplicación**: El examen de qué tipos específicos de propiedades aparecen duplicadas con mayor frecuencia puede revelar patrones significativos. Por ejemplo, podría haber mayor duplicación en ciertas zonas o tipos de propiedades.
- **Calidad de recolección de datos**: Una alta tasa de duplicados podría indicar problemas en los procesos de recolección, integración o gestión de datos inmobiliarios.

5. Conclusiones

El análisis realizado ha permitido identificar y cuantificar los registros duplicados en el conjunto de datos inmobiliarios de Medellín, cumpliendo con el objetivo principal establecido. Las principales conclusiones derivadas de este trabajo son:

1. **Metodología efectiva**: El enfoque implementado para la detección de duplicados, basado en la agrupación por todas las características y el conteo de frecuencias, proporciona un método efectivo y escalable para identificar redundancias en conjuntos de datos inmobiliarios.
2. **Necesidad de preprocesamiento**: El análisis confirma la importancia crítica de las etapas de limpieza y transformación de datos antes de cualquier análisis sustantivo. La estandarización de formatos numéricos, en particular, resultó esencial para la correcta identificación de duplicados.
3. **Conservación de información**: La estrategia de mantener registros con valores problemáticos (sustituyendo NaN por -1) permitió un análisis más completo, evitando la pérdida de información potencialmente valiosa.
4. **Base para análisis futuros**: El DataFrame resultante, que contiene registros únicos junto con su frecuencia, constituye una base sólida para análisis posteriores del mercado inmobiliario de Medellín, eliminando las distorsiones que podrían introducir los registros duplicados.

5. **Oportunidad de mejora en la gestión de datos**: La presencia de duplicados sugiere oportunidades de mejora en los procesos de recolección, validación y gestión de datos inmobiliarios para futuros estudios.