Instituto Nacional de Formación Técnica Profesional INFOTEP HVG

Valores Atípicos De Dataset houses_medellin

Robert López, Dayanna Montes, Santiago Orozco

Resumen

Este informe presenta un análisis de un archivo CSV con datos de viviendas en Medellín, utilizando Python y sus bibliotecas para identificar valores atípicos en cada columna. Se realizó una detección de anomalías, calculando la cantidad y el porcentaje de valores atípicos en el conjunto de datos. Además, se generó una visualización gráfica para facilitar la interpretación de los resultados.

1 Introducción

En el presente documento se mostrarán y se hará un análisis hecho a unos datos tomados de una encuesta hecha en la ciudad de Medellín (houses_medellin.csv). El objetivo será tomar este banco de datos y extraer todos los valores atípicos, el procedimiento lo haremos mediante la codificación. De esta forma visualizaremos los valores atípicos que contiene cada columna del banco de datos. Cabe mencionar que la definición de un valor atípico son aquellos datos que tienen un valor anormal. Los resultados finales se graficaran para dar un contexto más general del trabajo realizado.

2 Objetivos

2.1 Objetivo general

Extraer y analizar todos los valores atípicos que contienen cada columna haciendo un análisis estadístico para conocer el porcentaje de los datos anómalos.

2.2 Objetivos específicos

- Implementar bloques de código para analizar las columnas contenidas en un archivo de excel y graficar. El primero cumple la función de extraer valores atípicos y el segundo tiene como funcionalidad resumir y graficar los valores atípicos.
- Diseñar tablas para organizar los valores extraídos mediante la codificación.

- Explicar brevemente el porque de los valores atípicos representados en la gráfica.

3 Descripción de la actividad

Para dar inicio a este proceso reunimos algunas librerías y además de esto una pequeña investigación para saber como se utilizan cada una de ellas.

- Pandas: Librería para el manejo y análisis de datos en estructuras como DataFrames y Series.
- NumPy: Librería para cálculos numéricos y operaciones con arreglos/matrices.
- Matplotlib: Librería para la visualización de datos.
- Seaborn: Basada en Matplotlib que permite hacer gráficos más estilizados y estadísticos.
- String: Librería estándar de Python para manipulación de texto.

El trabajo de codificación se dividió en 2 partes tenemos el primer bloque tiene como objetivo extraer todos los valores atípicos de cada una de las columnas, aquí se aplica la fórmula del rango intercuartilico.

El **Rango Intercuartílico (IQR)** se define como la diferencia entre el tercer cuartil ($Q3$) y el primer cuartil ($Q1$):

$$IQR = Q3 - Q1$$

Para detectar valores atípicos, se usan los límites:

$$\text{Límite inferior} = Q1 - 1.5 \times IQR$$

$$\text{Límite superior} = Q3 + 1.5 \times IQR$$

Si un valor se encuentra fuera de estos límites, se considera un **outlier**.

Luego de esto pasamos a la segunda parte, aquí tenemos como finalidad generar una gráfica donde se muestre los valores atípicos y sus respectivos porcentajes, esta parte se complementa con el primer bloque de código y también se hace uso de las librería para que me genere la gráfica de detección de valores atípicos.

4 Habitaciones

Valores atípicos en Habitaciones
5.0
5.0
20.0
5.0
5.0
5.0

En la tabla se destaca un valor de 20 habitaciones, significativamente superior a los valores predominantes de 5 habitaciones. Esto sugiere que una vivienda con 20 habitaciones es un valor atípico, ya que no es común encontrar propiedades residenciales con tantas habitaciones. Este dato podría corresponder a un hotel, una residencia estudiantil.

5 Baños

Valores atípicos en Baños		
5.0	5.0	5.0
5.0	5.0	5.0
5.0	5.0	5.0
5.0	5.0	6.0
5.0	5.0	5.0
5.0	6.0	

En la tabla se identifican dos valores de 6.0, que son ligeramente superiores al valor predominante de 5.0. Esto sugiere que las viviendas con 5 o 6 baños no son muy comunes en la muestra, lo que podría indicar propiedades de mayor tamaño, de lujo o con un uso distinto al residencial estándar.

6 Estrato

Valores atípicos en Estrato
0

La variable estrato representada en misma nos deja analizar de que tiene un 0% de probabilidad de encontrar valor atípico debido a que en Colombia solo existen 6 estratos socioeconómicos establecidos.

7 Parqueaderos

Valores atípicos en Parqueaderos
5.0
5.0

Se analiza en la presente imagen la variable parqueaderos la cual nos indica 2 valores atípicos lo cual quiere decir que se encuentra en el conjunto de datos al menos 2 ocasiones en la que la variable parqueaderos contiene este valor anormal.

8 Precios

Valores atípicos de Precios			
1417743028	1850000000	1417743028	1850000000
1417743028	1200000000	1417743028	1417743028
15000000000	1500000000	1417743028	3000000000
1417743028	3000000000	1417743028	1200000000
1650000000	1417743028	1600000000	1648000000
1600000000	1417743028	1200000000	1350000000

La presente imagen representa la variable precio la cual nos muestra varios valores bastante inusual, podemos encontrar valores que se repiten (1,417,743,028) lo que puede ser un error en los datos y también vemos un valor atípico destacado (1,500,000,000,000) entre todos, lo que nos deja saber claramente que en Colombia no podemos encontrar una vivienda común por un precio tan exorbitante.

9 Área Construída (m^2)

Valores atípicos en Área construída (m^2)			
12597.0	8023.0	8678.0	297.0
257.0	235.0	250.0	8678.0
8023.0	297.0	257.0	235.0
250.0	12597.0	6139.0	248.0
5083.0	627.0	342.0	715.0
9966.0	1478.0	258.0	7654.0
342.0	368.0	330.0	

En la tabla se observan valores atípicos en el área construida, con grandes variaciones entre las propiedades. Algunos inmuebles presentan dimensiones extremadamente altas, como 12,597 m^2, 9,966 m^2 y 8,678 m^2, lo que sugiere que podrían tratarse de bodegas, edificios comerciales. Asimismo, se identifican valores atípicos en un rango más moderado, entre 235 m^2 y 715 m^2, que podrían corresponder a casas de gran tamaño.

10 Área Privada (m^2)

Valores atípicos en Área Privada (m^2)	
248.0	297.0
250.0	342.0
250.0	342.0
257.0	368.0
257.0	627.0
258.0	8023.0
297.0	8023.0

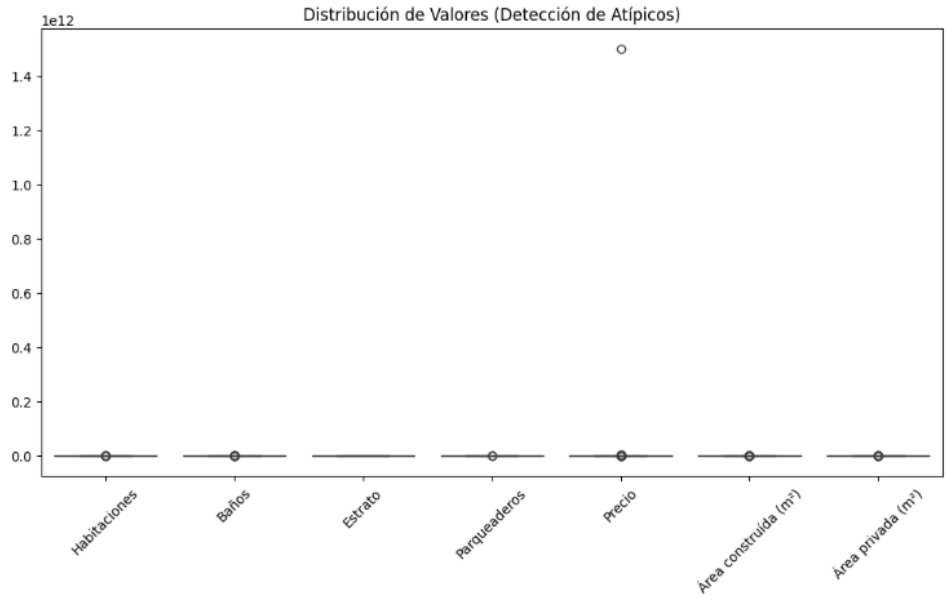
En la anterior tabla observamos los valores atípicos encontrados en el área privada, el valor más elevado que se encuentra fuera de lo normal es el 8023.0, probablemente este valor no corresponde a una vivienda común. Puede hacer referencia a una mansión, hotel, zona rural o un club. Pero no a una vivienda. Los valores como 250-627 pueden representar a apartamentos o casas muy grandes, aún así se considera un valor atípico o un área privada poco común.

11 Resumen de los valores atípicos

Variable	Total Valores Atípicos	Porcentaje de Datos Atípicos
Habitaciones	6.0	1.40%
Baños	17.0	3.95%
Estrato	0.0	0.00%
Parqueaderos	2.0	0.47%
Precio	24.0	5.53%
Área construida (m^2)	27.0	6.28%
Área privada (m^2)	14.0	3.26%

- Habitaciones: Para la variable habitaciones tenemos 6 valores que son atípicos con un porcentaje de 1.40%. Es decir que hay viviendas con bastantes habitaciones, algo fuera de lo normal. Al menos que los datos hagan referencia a otro tipo de propiedad.
- Baños: Para la variable baños encontramos 17 valores atípicos lo que nos deja un porcentaje de 3.95% lo que quiere decir que hay viviendas o propiedades con un alto número de baños, algo que es inusual.
- Estrato: En la presente variable encontramos el porcentaje total de la variable estrato este porcentaje fue extraído de 430 datos, donde solo 309 nos dice estadísticamente que hay 0% de encontrar un nivel socioeconómico más allá de lo que esta establecido.
- Parqueaderos: Para la variable parqueaderos tenemos un porcentaje de 0.47% de 430 datos recolectados, aquí encontramos solo 2 ocasiones en la que se encuentra 2 valores atípicos. Tal cual como se muestra aquí.
- Precio: En la variable precio se encontraron 24 casos de valores atípicos lo que nos deja un porcentaje de 5.53% algo inusual, lo que nos indica que pueden haber datos erróneos dentro de la estadística.
- Área Construída (m^2): En esta variable encontramos 27 valores atípicos, observamos que hay valores que son anómalos, muy altos y poco comunes. El porcentaje equivalente a los valores atípicos encontrados en área construída es de 6.28%

- Área Privada (m^2): En esta variable encontramos 14 datos anómalos, equivalente a un 3.26% algo inusual y poco común, también puede indicar datos erróneos dentro de la estadística.



La presente gráfica muestra la totalidad de valores atípicos encontrados en cada una de las variables. Visualizamos en la imagen que en el eje X tenemos las variables por rangos (bigotes) lo que nos indica los datos típicos del análisis realizado y encontramos pequeños círculos en cada uno de estos rangos lo que quiere decir que representa los valores atípicos hallados en el análisis de los datos.

Por otro lado, se visualiza en el eje Y el porcentaje de cada valor atípico encontrado en los datos, vemos que de las 7 variables analizadas solo la variable precio sobrepasa el 1.4%, mientras las otras variables no pasan de 0.0% ya que son valores atípicos muy bajos dentro de los datos recolectados. La información que se muestra en la gráfica se confirma usando la fórmula de conversión de porcentaje a fracción. Se utiliza para expresar un porcentaje como una fracción dividiendo el número entre 100. En este caso:

$$13\% = \frac{13}{100}$$

También se puede expresar en forma decimal:

$$13\% = \frac{13}{100} = 0.13$$

12 Conclusión

A través de este trabajo se pudo analizar los valores aípicos de diferentes variables como: Habitaciones, baños, precio, parqueaderos, área construída, estrato y área privada. Estos valores atípicos o anómalos pueden deberse a errores en la recolección de datos, inconsistencias en el registro de la información o a la existencia de propiedades con características poco comunes. El uso del rango intercuartílico (IQR) permitió detectar estos valores de manera efectiva, facilitando la interpretación y visualización de los datos.
En conclusión, la detección de outliers o valores atípicos en los análisis de datos es importante para garantizar una mejor toma de decisiones y mejorar la calidad de los resultados obtenidos.