Estudio de Variables Mediante Técnicas Estadisticas

Luis Correa, Valeria Márquez, Maily Silva

Resumen

En esta investigación, se realizó un análisis de datos utilizando metodos estadisticos, tales como: análisis univariado y análisis bivariado, para estudiar siete variables las cuales son: housing-median-age, total-bedrooms, population, households, median-income y median-house-valuie. En donde se emplearon histogramas, medidas de centralidad y medidas de dispersión para obtener una visión detallada de la distribución y caracteristicas de cada variable. A partir de estos análisis, se generaron veintíun gráficas, que proporcionan una comprensión clara y profunda de las propiedades individuales de cada variable. Tambien, re realizó un análisis bivariado usando tres variables: housing-median-age, total-rooms y median-house-value, con el fin de identificar posibes patrones de correlación Para este análisis, se crearon gráficas especificas que permitieron una exploración más detallada de las relaciones entre estas variable.

Palabras Claves: análisis univariado, análisis bivariado, datos, medidas de centralidad, variables.

1 Introducción

El análisis de variables es una parte fundamental de la estadística, ya que permite comprender, interpretar y extraer conclusiones a partir de los datos. A través de diferentes técnicas estadísticas, es posible identificar patrones, relaciones y tendencias que facilitan la toma de decisiones en diversos campos. En esta investigación, se realizó un análisis univariado y bivariado de un conjunto de datos que contiene información sobre viviendas.

El análisis univariado hizo un enfoque en la distribución y las caracteristicas estadísticas de las variables individualmente, permitiendo identificar tendencias, valores atípicos y rangos de variabilidad. Por otro lado, el análisis bivariado exploró la relación entre variables claves para identificar posibles patrones de correlación y dependencia, lo que puede proporcionar información valiosa sobre el comportamiento. De tal manera que pudimos extraer conclusiones significativas que nos ayudó a una mejor interpretación de los datos analizados.

2 Objetivos

- Objetivo General

  Realizar un estudio de variables mediante la aplicación de técnicas estadísticas, con el propósito de analizar su comportamiento, identificar

patrones y establecer relaciones significativas entre los datos. Para ello, se empleará el análisis univariable y el análisis bivariado, complementados con la generación de gráficos estadísticos que faciliten la visualización y comprensión de la información. Este análisis permitirá extraer conclusiones relevantes y respaldar la toma de decisiones en distintos ámbitos del conocimiento.

- Objetivos Especificos

  1. Analizar la distribución y las características estadísticas de las variables del conjunto de datos.
  2. Representar gráficamente la información para facilitar su interpretación.
  3. Identificar patrones y tendencias en los datos a través del análisis univariado y bivariado.
  4. Explorar la relación entre variables clave y evaluar su grado de correlación.

3 Descripción de los análisis

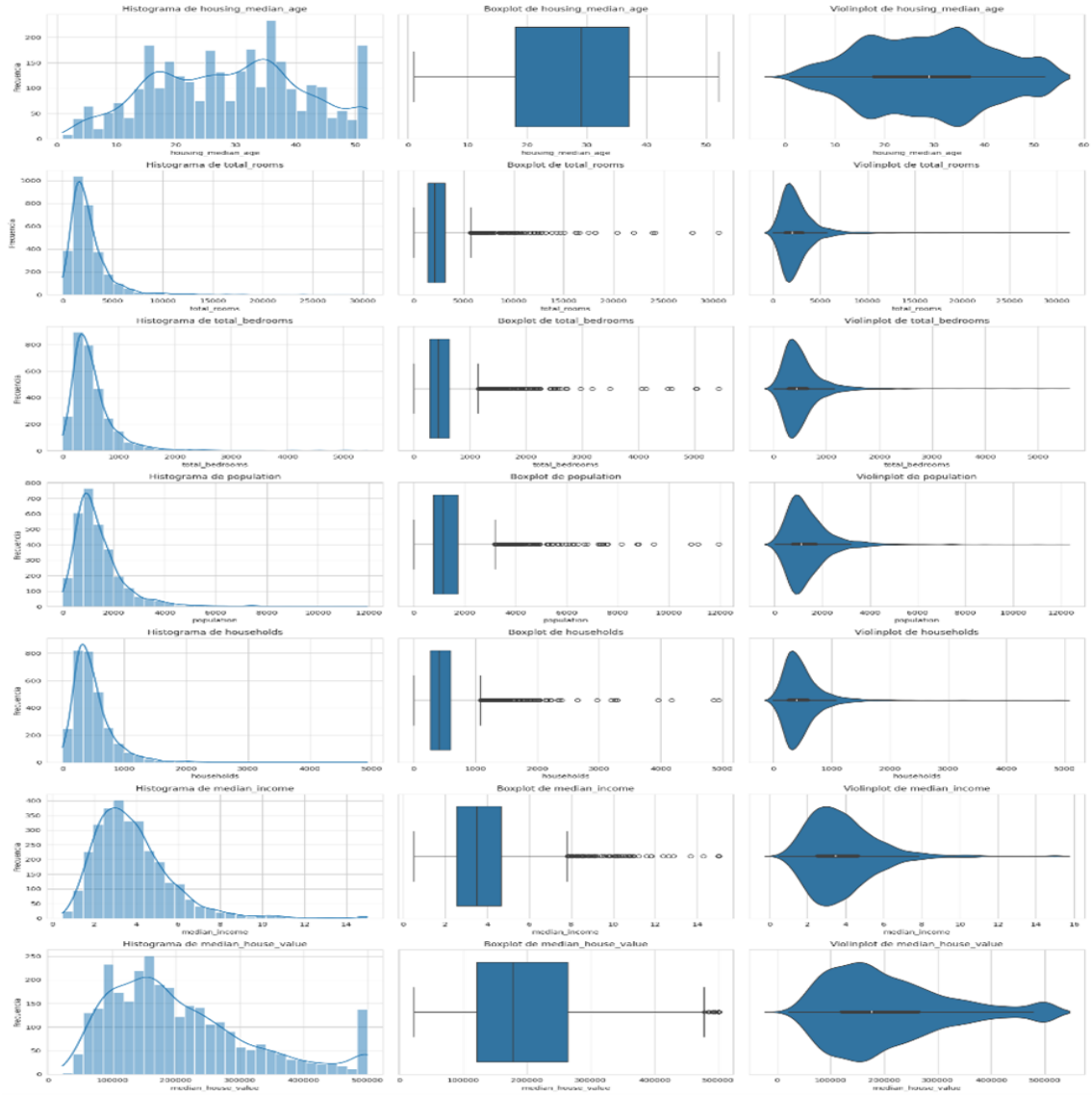
- Análisis Univariado
  - El análisis univariado se centró en estudiar las características individuales de siete variables del conjunto de datos: housing-median-age, total-rooms, total-bedrooms, population, households, median-income y median-house-value. Para cada una de estas variables, se utilizaron tres tipos de gráficas: histogramas, boxplots y violinplots. Los histogramas ayudaron a observar la distribución de los datos, identificando posibles sesgos o picos de frecuencia. Los boxplots proporcionaron información sobre la mediana, la dispersión y los posibles valores atípicos. Por último, los violinplots ofrecieron una visualización detallada de la densidad de la distribución y cómo los valores se dispersan alrededor de la mediana, permitiendo una comprensión más profunda de la variabilidad de cada variable.

- Análisis Bivariado
  - El análisis bivariado se enfocó en explorar las relaciones entre tres variables clave: housing-median-age, total-rooms y median-house-value. Para ello, se utilizaron gráficas de dispersión para visualizar cómo las variables interactúan entre sí y si existe alguna relación lineal o no lineal. Además, se construyó una matriz de correlación que cuantificó el grado de relación entre las variables, ayudando a identificar si las variables están correlacionadas de manera positiva o negativa. Este análisis permitió descubrir las relaciones clave entre las variables,

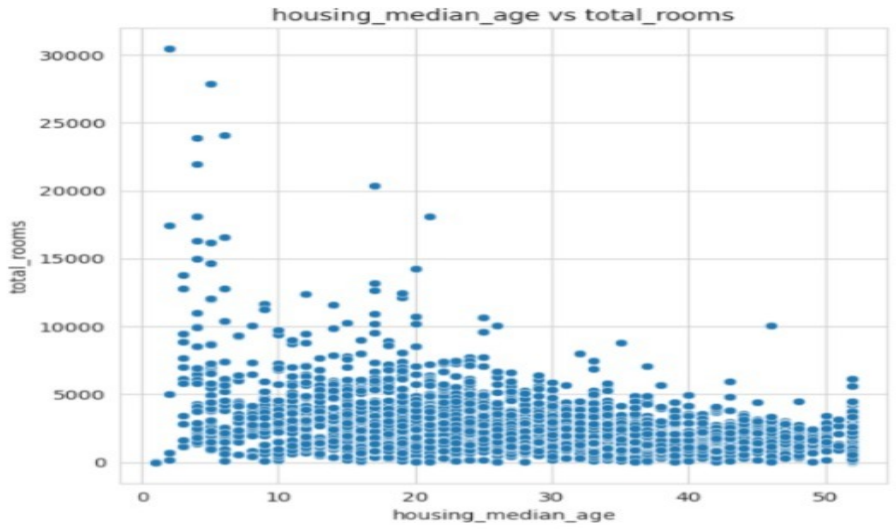
proporcionando una base sólida para futuras investigaciones sobre los factores que influyen en los precios de las viviendas.

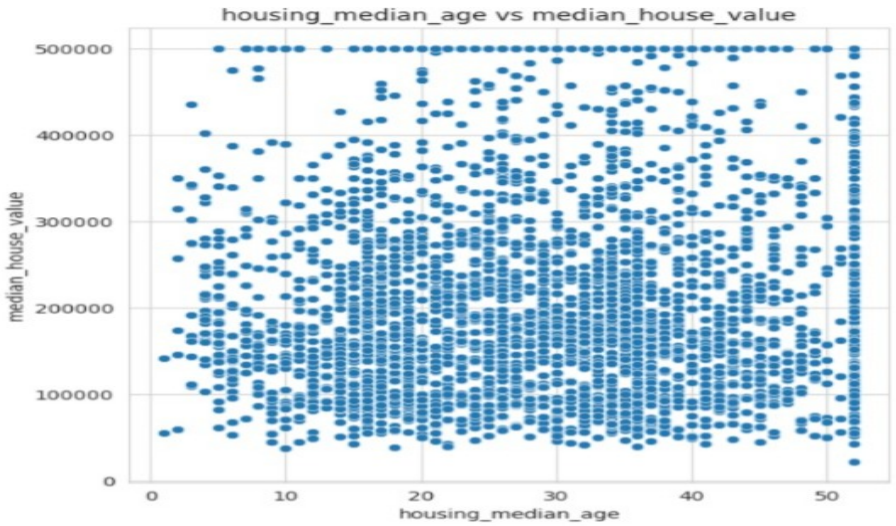
4 Gráficas de Resultados

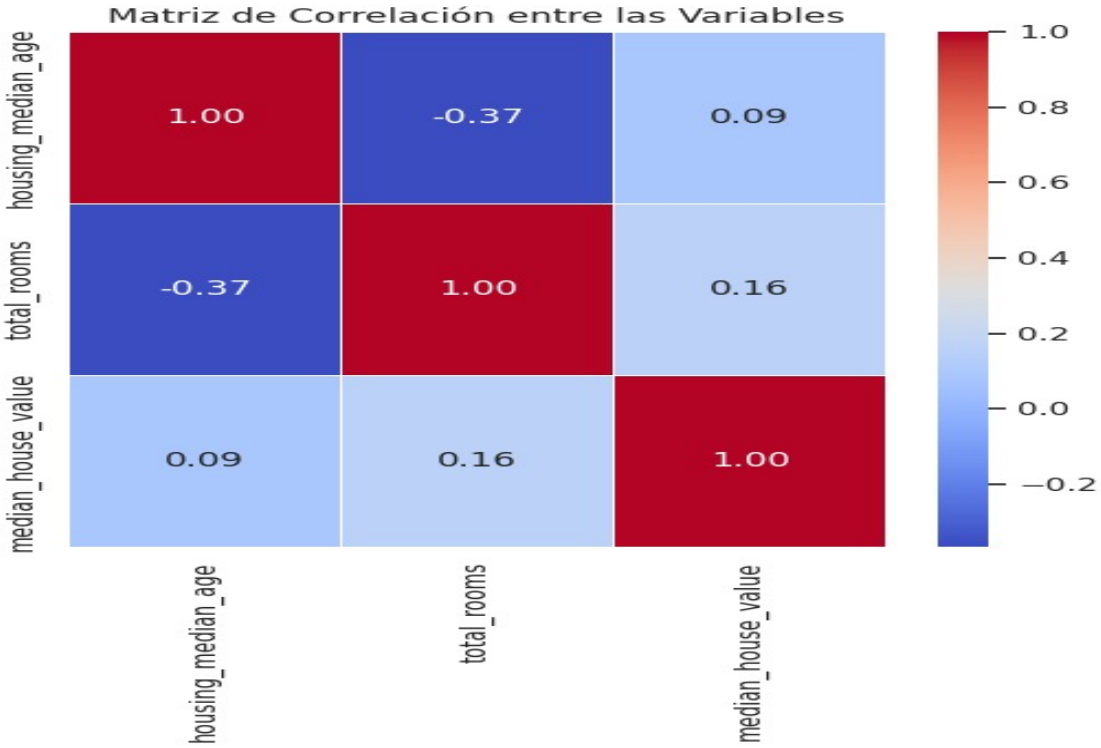
- **Análisis Univariado**



- **Análisis Bivariado**




housing_median_age vs median_house_value
median_house_value
housing_median_age
500000
400000
300000
200000
100000
0
0
10
20
30
40
50


Matriz de Correlación entre las Variables
housing_median_age
total_rooms
median_house_value
1.00
-0.37
0.09
-0.37
1.00
0.16
0.09
0.16
1.00
housing_median_age
total_rooms
median_house_value
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

5 Análisis de cada Gráfica

5.1 Análisis Univariado

- Para la Variable housing-median-age

    1. Histograma: La distribución parece bimodal, con picos alrededor de 15 y 50 años. Esto indica que hay dos grupos predominantes de viviendas: unas más recientes y otras más antiguas.
    2. Boxplot: No se observan valores atípicos significativos. La mediana está en torno a los 30-35 años.
    3. Violinplot: Confirma la distribución bimodal, con mayor densidad en las edades mencionadas.

- Para la Variable total-rooms

    1. Histograma: Distribución asimétrica positiva (sesgada a la derecha). La mayoría de las viviendas tienen menos de 5000 habitaciones, pero hay valores extremos hasta 30,000.
    2. Boxplot: Muestra muchos valores atípicos, indicando la existencia de viviendas con un número inusualmente alto de habitaciones.
    3. Violinplot: Muestra que la mayoría de los datos están concentrados en el extremo inferior, con una larga cola hacia valores más altos.

- Para la Variable total-bedrooms

    1. Histograma: Similar al de total-rooms, con una distribución sesgada a la derecha y una gran cantidad de viviendas con menos de 1000 dormitorios.
    2. Boxplot: Muestra valores atípicos en viviendas con más de 2000 dormitorios.
    3. Violinplot: La densidad más alta está en valores bajos, pero con una cola larga hacia valores más altos.

- Para la Variable population

    1. Histograma: Distribución sesgada a la derecha, con la mayoría de las zonas teniendo menos de 2000 habitantes.
    2. Boxplot: Existen muchos valores atípicos en las zonas más densamente pobladas.
    3. Violinplot: Se observa una alta concentración en poblaciones bajas y una cola alargada hacia valores altos.

- Para la Variable households

1. Histograma: Distribución similar a la de la población, sesgada a la derecha, con la mayoría de las zonas teniendo menos de 1000 hogares.
2. Boxplot: Presenta valores atípicos para zonas con más de 2000 hogares.
3. Violinplot: La densidad es más alta en los valores bajos, con una distribución en forma de "gota" similar a la población.

- Para la Variable median-income

1. Histograma: Distribución sesgada a la derecha, con la mayoría de los ingresos en el rango de 2 a 6.
2. Boxplot: Hay valores atípicos en ingresos altos (mayores a 12).
3. Violinplot: Muestra una mayor densidad en ingresos bajos y una caída en ingresos altos.

- Para la Variable median-house-value

1. Histograma: Distribución sesgada a la derecha, con valores frecuentes entre 100,000 y 300,000, pero con un pico en 500,000, lo que sugiere un límite en los datos.
2. Boxplot: Se observan valores atípicos en viviendas de alto valor.
3. Violinplot: Confirma la alta densidad en valores bajos y una caída en valores altos, con un límite en 500,000.

5.2 Análisis Bivariado

- Para las Variables housing-median-age vs. total-rooms

1. Se observa una gran concentración de puntos en la parte inferior izquierda del gráfico.
2. A medida que la antigüedad de las viviendas (housing-median-age) aumenta, el número total de habitaciones (total-rooms) tiende a disminuir.
3. Hay algunos valores extremos con muchas habitaciones en viviendas más nuevas, pero no parece haber una relación lineal fuerte.
4. No hay una correlación clara entre la antigüedad de las viviendas y el número de habitaciones.

- Para las Variables housing-median-age vs. median-house-value

1. No se observa una tendencia clara entre la edad mediana de las viviendas (housing-median-age) y el valor de las casas (median-house-value).
2. Se nota una acumulación de puntos en la parte superior, indicando que muchas casas alcanzan un valor máximo

3. Hay variabilidad en los precios para todas las edades de vivienda, lo que indica que la antigüedad no es un factor determinante en el valor de la casa.
4. La edad de las viviendas no parece afectar significativamente su valor, pero el límite de precios podría estar ocultando tendencias.

- Para las Variables total-rooms vs. median-house-value

1. La mayoría de los puntos están concentrados en la parte izquierda del gráfico, lo que sugiere que muchas casas tienen relativamente pocas habitaciones.
2. No parece haber una tendencia clara que indique que más habitaciones significan un mayor valor de la vivienda.
3. hay valores extremos donde algunas viviendas con muchas habitaciones tienen tanto valores bajos como altos.
4. El número total de habitaciones por sí solo no determina el valor de la vivienda. Puede ser más relevante analizar la cantidad de habitaciones por unidad habitacional en lugar del total absoluto.

- Para la matriz de correlación

1. La diagonal Principal Siempre es 1, porque cualquier variable está perfectamente correlacionada consigo misma.
2. housing-median-age vs. total-rooms: Correlación de -0.37, lo que indica una relación inversa moderada (a medida que la edad media de las viviendas aumenta, la cantidad total de habitaciones tiende a disminuir).
3. housing-median-age vs. median-house-value: Correlación de 0.09, lo que sugiere una relación casi nula (la edad de las viviendas no influye mucho en el valor medio de la casa).
4. total-rooms vs. median-house-value: Correlación de 0.16, lo que indica una relación positiva débil (a más habitaciones, el valor medio de la casa tiende a aumentar ligeramente).

El mapa de calor usa colores para representar los valores de correlación:
- Rojo intenso (1.0 o cercano a 1.0): Correlación positiva fuerte.
- Azul intenso (cercano a -1.0): Correlación negativa fuerte.
- Tonos claros: Correlaciones débiles o inexistentes.

6 Conclusión

El análisis de datos realizado permitió conocer la distribución y características estadísticas de las variables incluidas este estudio. A través del análisis univariado,

se identificaron las tendencias centrales y la dispersión de cada variable, lo que facilita una comprensión más profunda del conjunto de datos. En cambio, el análisis bivariado reveló posibles correlaciones entre las variables estudiadas, proporcionando información valiosa para tener una visión mas clara. La combinación de estos análisis demuestra la importancia de aplicar técnicas estadísticas en la exploración y visualización de datos para obtener información significativa y fundamentada.