Taller Analisis Univariado y Bivariado

Sebastián Cotes, Yefferson González, Lorann Peñuela

IES, INFOTEP Instituto Nacional de Formación Técnica Profesional "HVG"

Humberto Velazquez Garcia

Abstract

En esta actividad, se nos pide analizar los datos del archivo "california housing test.csv" enla plataforma Google CodeLab. Utilizaremos librerías como matplotlib y numpy. Para leer los datos desde un archivo CSV y calcurar la media, mediana, desviación estándar y varianza, para varias variables univariadas. Ademas, la matriz de correlación para un conjunto de variables bivariadas y su respectivo gráficos para visualizar las relaciones entre estas variables. Este proceso nos permite analizar y visualizar los datos de manera efectiva, proporcionando una comprensión más profunda de las características y relaciones dentro del conjunto de datos.

1 Introducción

El análisis de datos es una herramienta fundamental para comprender y tomar decisiones informadas basadas en grandes volúmenes de información. Para poder analizar esta información, utilizaremos el lenguaje de programación Python y algunas librerías disponibles en el gestor Google CodeLab, esto con el objetivo de ir mejorando y comprendiendo como se identifican patrones y relaciones entre diferentes variables, utilizando los lenguajes de programación.

2 Objetivos

Los objetivos principales de esta actividad son:

- Realizar una análisis univariado, para las variables: housing median age, total rooms, total bedrooms, population, households, median income, median house value.
- Realizar un análisis bivariado entre las variables: housing median age, total rooms, median house value.
- Genera las gráficas necesarias y realizar un análisis de los resultados.

3 Descripción de los Análisis

Se realizaran análisis univariados y bivariados utilizando las librerías pandas y matplotlib en Python. Los análisis univariados incluyeron el cálculo de la media,

mediana, desviación estándar y varianza, para cada variable en el conjunto de datos. Los análisis bivariados incluyen el cálculo de la matriz de correlación y la generación de gráficos de dispersión para visualizar las relaciones entre variables.

Utilizaremos pandas para leer los datos de ejemplo proporcionados por Google colab. Pandas, nos permite cargar los datos en un DataFrame, que es una estructura de datos muy eficiente para manipular y analizar datos tabulares. Ademas nos permite realizar los cálculos estadísticos y el análisis de correlación [1].

Por otra parte encontramos a matplotlib. Matplotlib nos permitirá crear gráficos, generar histogramas y crear los gráficos de dispersión para visualizar la distribución y las relaciones entre las variables [2].

4 Conclusiones

El análisis realizado proporciona una visión detallada de las características y relaciones entre las variables del conjunto de datos de viviendas en California. Los histogramas revelan la distribución de cada variable, mientras que los gráficos de dispersión ayudan a identificar posibles correlaciones. Estos resultados pueden ser útiles para futuras investigaciones y toma de decisiones en el ámbito inmobiliario. En general, se observa que la mayoría de las viviendas son relativamente nuevas y tienen un número moderado de habitaciones. Además, existe una ligera correlación positiva entre la edad de las viviendas y su valor mediano, lo que sugiere que las viviendas más antiguas pueden estar ubicadas en áreas más deseables.

5 Gráficas de Resultados

Terminada la **codificación**, y teniendo en cuenta los requerimientos iniciales, obtenemos como resultado las siguientes gráficas.

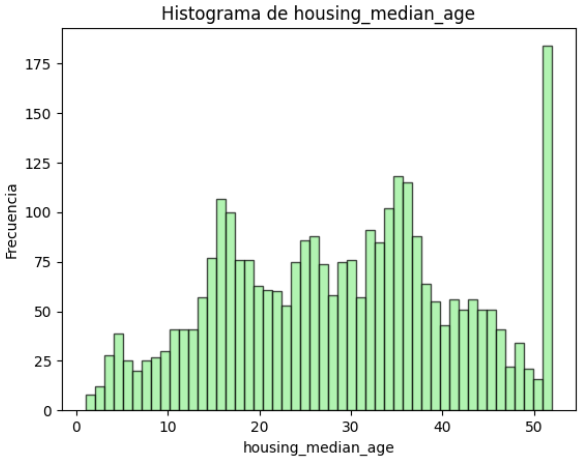

Figure 1: **housing-median-age**
La mayoría de las viviendas tienen una edad mediana entre 15 y 30 años, con un pico notable alrededor de los 20 años. Esto sugiere que muchas viviendas en el conjunto de datos son relativamente nuevas.

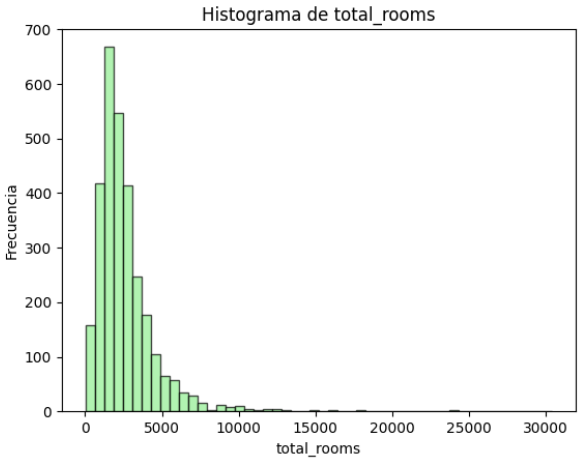

Figure 2: **total-rooms**
La distribución es sesgada a la derecha, con la mayoría de las viviendas teniendo entre 2,000 y 4,000 habitaciones. Las viviendas con un número muy alto de habitaciones podrían ser propiedades más grandes o edificios multifamiliares.

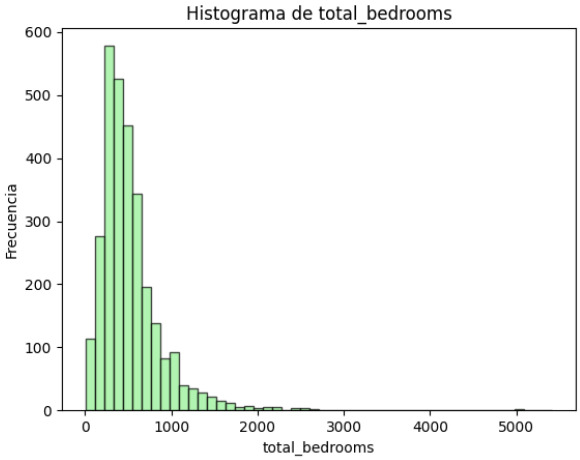

Figure 3: **total-bedrooms**
La mayoría de las viviendas tienen entre 500 y 1,000 dormitorios. Las viviendas con un número muy alto de dormitorios podrían ser propiedades grandes o multifamiliares.

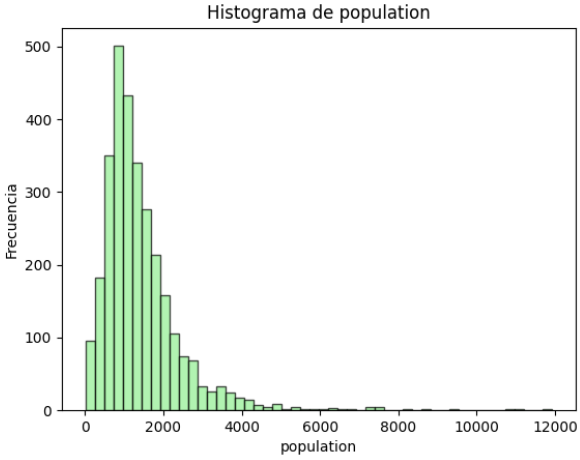

Figure 4: **population**
La mayoría de las áreas tienen una población entre 1,000 y 3,000 personas. Las áreas con poblaciones muy altas podrían ser zonas urbanas densamente pobladas.

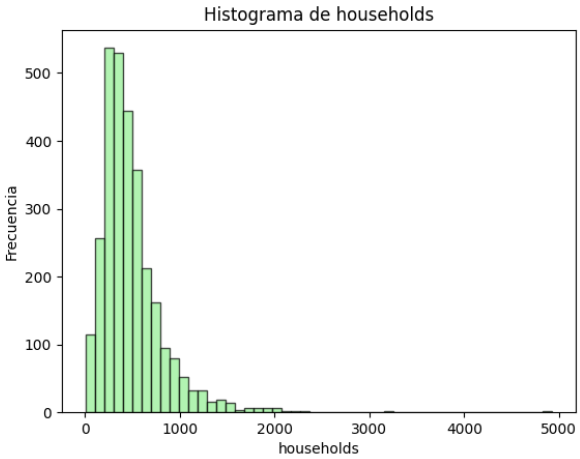

Figure 5: **households**
La mayoría de las áreas tienen entre 500 y 1,500 hogares. La distribución es similar a la de la población.

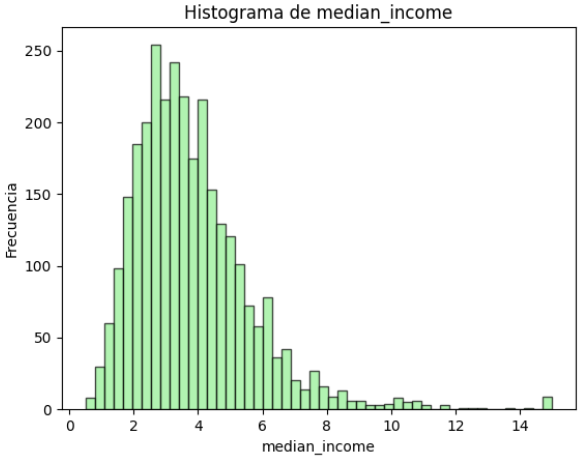

Figure 6: **median-income**
La mayoría de los ingresos medianos están entre 2 y 6 (en decenas de miles). Esto indica una distribución relativamente uniforme con una ligera tendencia hacia ingresos medianos más bajos.

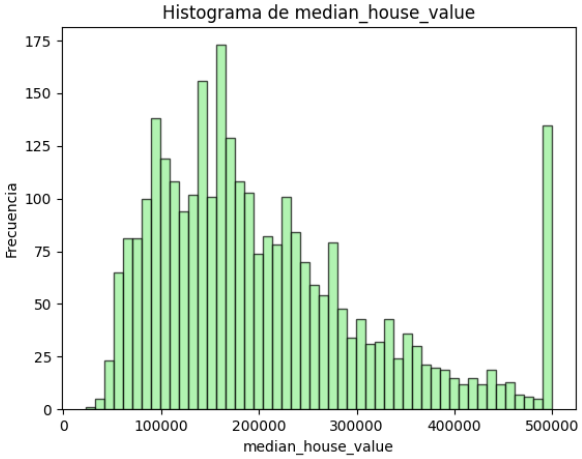

Figure 7: **median-house-value**
La mayoría de los valores medianos están entre 100,000 y 300,000 dólares. Las viviendas con valores muy altos podrían ser propiedades de lujo.

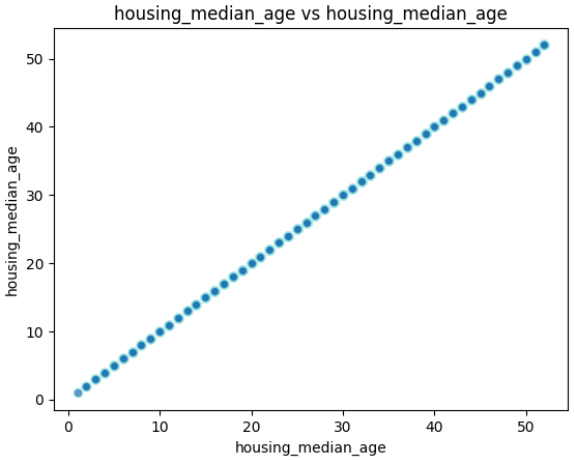

Figure 8: **housing-median-age-vs-housing-median-age**

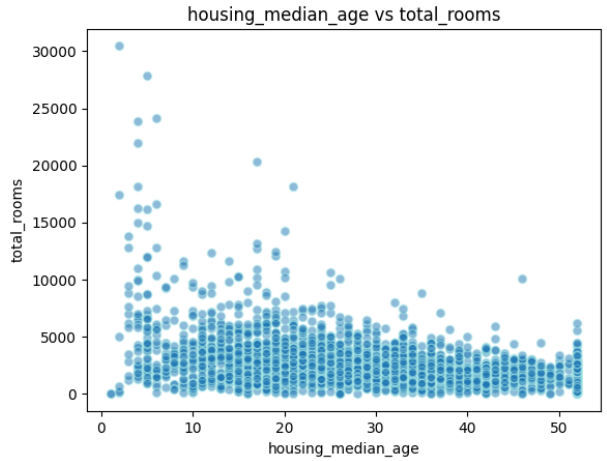

Figure 9: **housing-median-age-vs-total-rooms**
Se evidencia que la edad de las viviendas no influye significativamente en el número de habitaciones que tienen.

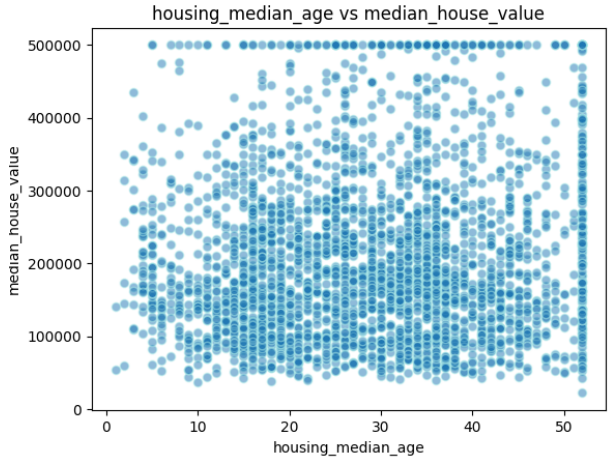

Figure 10: **housing-median-age-vs-median-house-value**

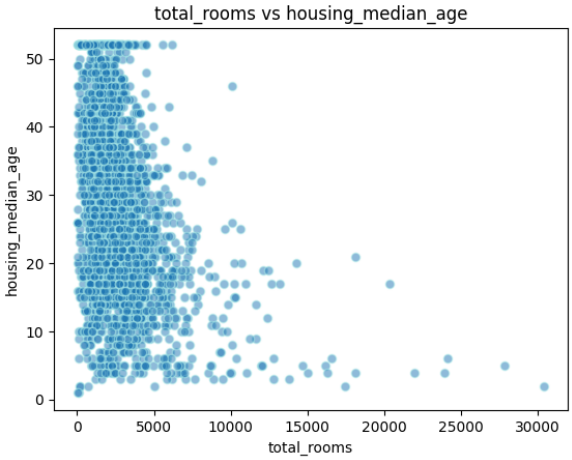

Figure 11: **total-rooms-vs-housing-median-age**
El valor de las viviendas no está directamente relacionado con su tamaño en términos de habitaciones.

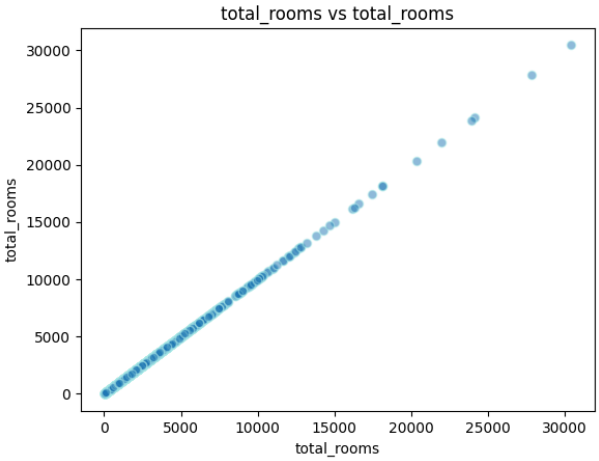

Figure 12: **total-rooms-vs-total-rooms**

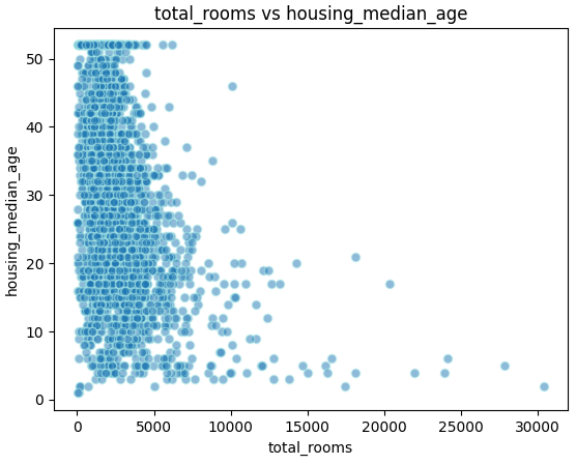

Figure 13: **total-rooms-vs-housing-median-age**
El tamaño de las viviendas, en términos de habitaciones, no está relacionado con su antigüedad.

Figure 14: **median-house-value-vs-housing-median-age**
Se muestra una ligera correlación positiva entre la edad de las viviendas y su valor mediano. Esto sugiere que las viviendas más antiguas pueden tener un valor más alto debido a su ubicación o características históricas.

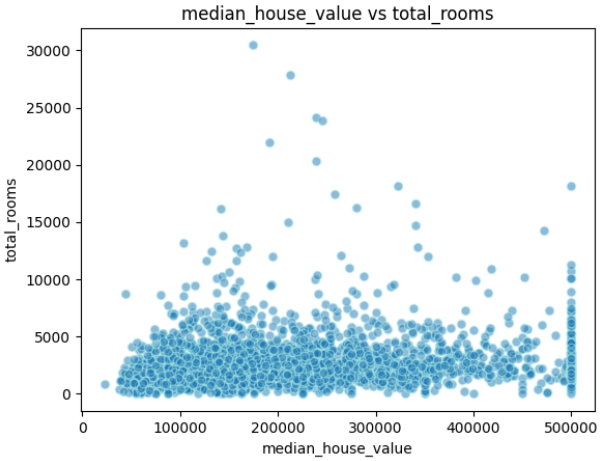

Figure 15: **median-house-value-vs-total-rooms**
El valor de las viviendas no está directamente relacionado con su tamaño en términos de habitaciones.

References

[1] DataScientest, Pandas : La biblioteca de Python dedicada a la Data Science [online], 10 2023. available from: `https://datascientest.com/es/pandas-python`.

[2] given=Daniel given i=D, Matplotlib: todo lo que tienes que saber sobre la librería python de dataviz [online], available from: `https://datascientest.com/es/todo-sobre-matplotlib`.