Análisis del Conjunto de Datos California Housing*

Critian Orozco, Marlon Caviedes, Alan Miranda

Infotep

Abstract

Este informe expone un estudio exploratorio del conjunto de datos *california_housing_test.csv*. Se efectúa un análisis univariado de las variables `housing_median_age`, `total_rooms`, `total_bedrooms`, `households`, `median_income` y `median_house_value` mediante histogramas y el cálculo de estadísticas básicas (media, mediana, moda, desviación estándar, varianza, rango e IQR). Asimismo, se examina la relación entre `median_house_value`, `total_rooms` y `total_bedrooms` usando gráficos de dispersión y mapas de calor. Los hallazgos ayudan a comprender tanto la distribución individual como la interrelación entre las variables.

Keywords: California Housing, Análisis Univariado, Análisis Bivariado, Histogramas, Correlación.

1 Introducción

El dataset *california_housing_test.csv* se utiliza comúnmente para aprender técnicas de análisis de datos y modelado. En este informe se estudian aspectos individuales y conjuntos de variables relacionadas con características de viviendas en California, lo cual resulta útil para detectar tendencias y relaciones en el mercado inmobiliario.

2 Objetivos

Los fines de este estudio son:

- Realizar un análisis univariado de las variables seleccionadas mediante histogramas y medidas estadísticas básicas.
- Investigar la relación entre `median_house_value`, `total_rooms` y `total_bedrooms` para detectar correlaciones.

3 Metodología

Se emplearon las librerías `pandas`, `matplotlib` y `seaborn` de Python para cargar los datos y generar los gráficos. En el análisis univariado se construyeron histogramas y se calcularon medidas de tendencia central y dispersión. Para el análisis

*Este trabajo forma parte de las actividades del curso de Ingeniería Informática.

bivariado se realizaron gráficos de dispersión (pairplot) y se obtuvo la matriz de correlación, la cual se visualizó con un mapa de calor. A continuación, se muestra un fragmento del código utilizado:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar el dataset
df = pd.read_csv('/content/sample_data/california_housing_test.csv')

# Análisis univariado: histogramas y estadísticas
for var in ['housing_median_age', 'total_rooms', 'total_bedrooms',
            'households', 'median_income', 'median_house_value']:
    plt.figure(figsize=(8,4))
    plt.hist(df[var].dropna(), bins=30, color='blue', edgecolor='black')
    plt.title('Histograma de ' + var)
    plt.xlabel(var)
    plt.ylabel('Frecuencia')
    plt.show()

# Análisis bivariado: pairplot y mapa de calor
sns.pairplot(df[['median_house_value', 'total_rooms', 'total_bedrooms']])
plt.suptitle('Dispersión entre variables', y=1.02)
plt.show()

corr_matrix = df[['median_house_value', 'total_rooms', 'total_bedrooms']].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Mapa de calor de correlaciones')
plt.show()
```

4 Resultados

Se generaron los siguientes gráficos:

5 Análisis e Interpretación

5.1 Análisis Univariado

La Figura 1 ilustra la distribución de la variable `housing_median_age`. Se observa que la mayoría de las viviendas tienen edades concentradas en un rango específico. Las medidas de tendencia central y dispersión permiten identificar el comportamiento típico y la variabilidad de la distribución.

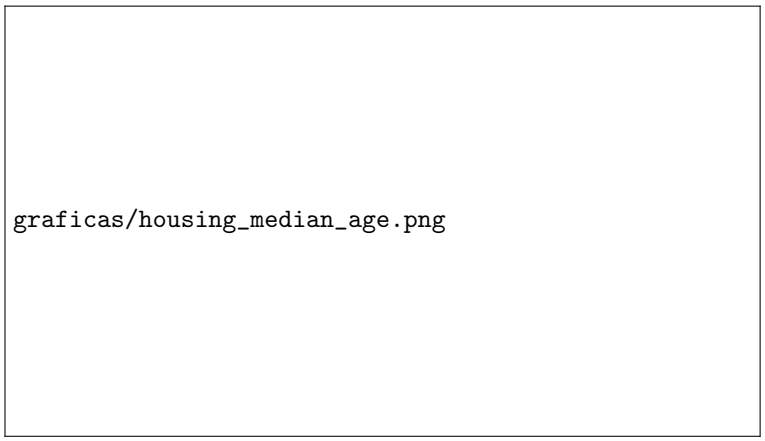

Figure 1: Histograma de `housing_median_age`.

5.2 Análisis Bivariado

Los gráficos de dispersión (Figura 2) muestran relaciones claras entre el valor mediano de las viviendas y el número de habitaciones, sugiriendo una correlación positiva. Asimismo, el mapa de calor (Figura 3) confirma una alta correlación entre `total_rooms` y `total_bedrooms`, lo cual respalda la hipótesis de que, a mayor cantidad de habitaciones, se incrementa el número de dormitorios.

6 Conclusiones

El estudio realizado permite concluir que:

- El análisis univariado revela la distribución y dispersión de cada variable, facilitando la comprensión de la estructura del dataset.
- El análisis bivariado indica relaciones significativas entre el tamaño de las viviendas y su valor, lo que puede servir como base para modelos predictivos.

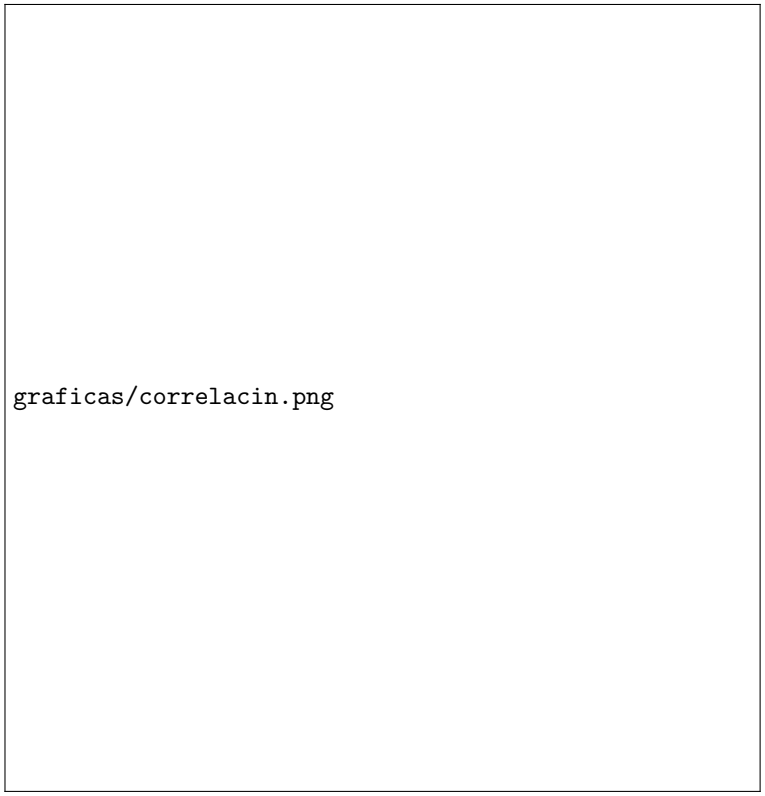

Figure 2: Gráficos de dispersión para `median_house_value`, `total_rooms` y `total_bedrooms`.

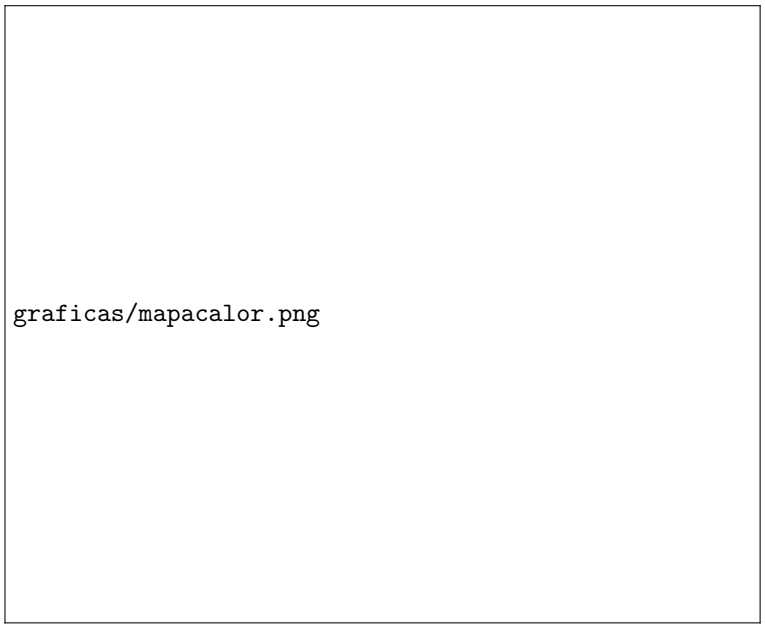

Figure 3: Mapa de calor de la matriz de correlación.