Introducción a la Inteligencia Artificial
Trabajo Práctico 2

Nombre: Eduardo Echeverria (a1516)

Se requiere construir una regresión que nos permita predecir el valor medio de las casas en distritos de California, EEUU (medidos en cientos de miles de dólares $100,000). Este dataset se deriva del censo de 1990 de EEUU, donde cada observación es un bloque. Un bloque es la unidad geográfica más pequeña para la cual la Oficina del Censo de EEUU publica datos de muestra (un bloque típicamente tiene una población de 600 a 3000 personas).

Los atributos, en el orden en que se guardaron en el dataset, son:

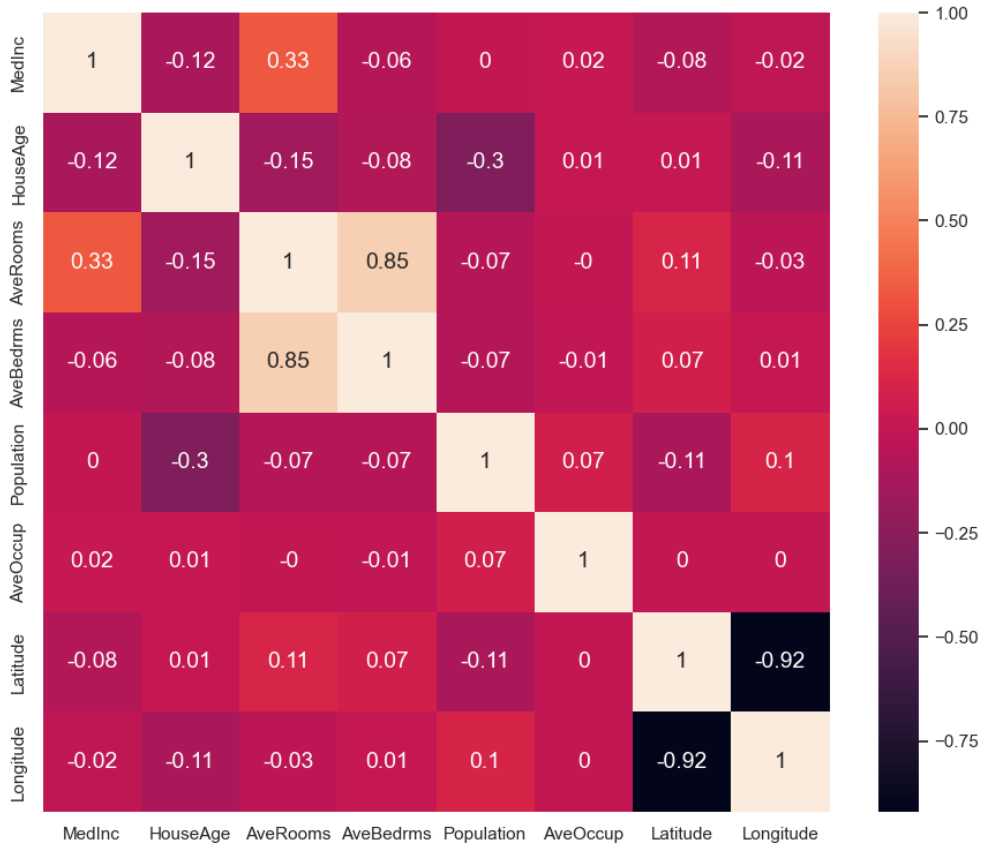
- MedInc: Ingreso medio en el bloque
- HouseAge: Edad mediana de las casas en el bloque
- AveRooms: Número promedio de habitaciones por hogar.
- AveBedrms: Número promedio de dormitorios por hogar.
- Population: Población del bloque
- AveOccup: Número promedio de miembros por hogar.
- Latitude: Latitud del bloque
- Longitude: Longitud del bloque

Y el target es:

- MedHouseVal: Mediana del costo de casas en el bloque (en unidades de a $100.000)

1. Obtener la correlación entre los atributos y los atributos con el target. ¿Cuál atributo tiene mayor correlación lineal con el target y cuáles atributos parecen estar más correlacionados entre sí? Se puede obtener los valores o directamente graficar usando un mapa de calor.

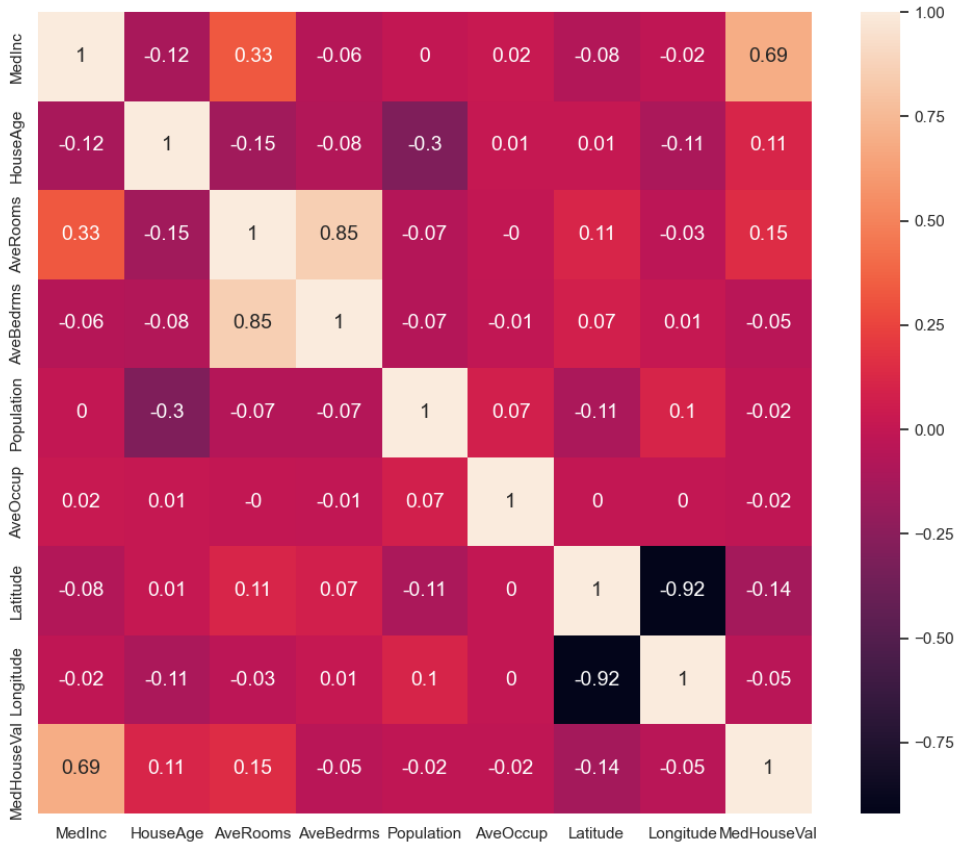
Correlación entre los atributos:



En este gráfico se puede observar que los atributos que tienen mayor correlación, con un valor de 0.85, son el número promedio de habitaciones por hogar (AveRooms) y el número promedio de dormitorios por hogar (AveBedrms).

También se observa una correlación importante (aunque menor que la anterior) con un valor de 0.33 entre el promedio de habitaciones por hogar (AveRooms) y el ingreso medio en el bloque (MedInc)

Correlación entre los atributos con el target:

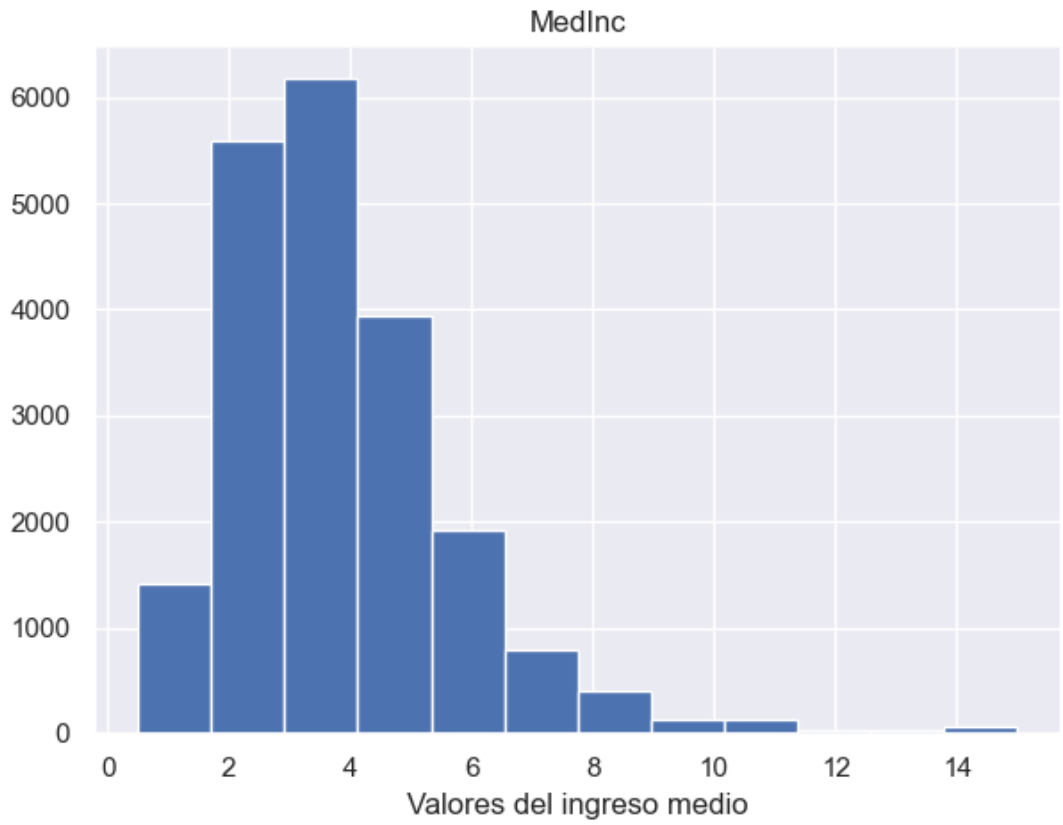


En este gráfico, además de las correlaciones mencionadas arriba, se observa también un valor de correlación alto de 0.69 entre el ingreso medio en el bloque (MedInc) y el target, es decir con la mediana del costo de las casas del bloque (MedHouseVal).

En este sentido, el atributo de ingreso medio en el bloque (MedInc) es el que tiene mayor correlación lineal con el target, es decir con la mediana del costo de las casas del bloque (MedHouseVal).

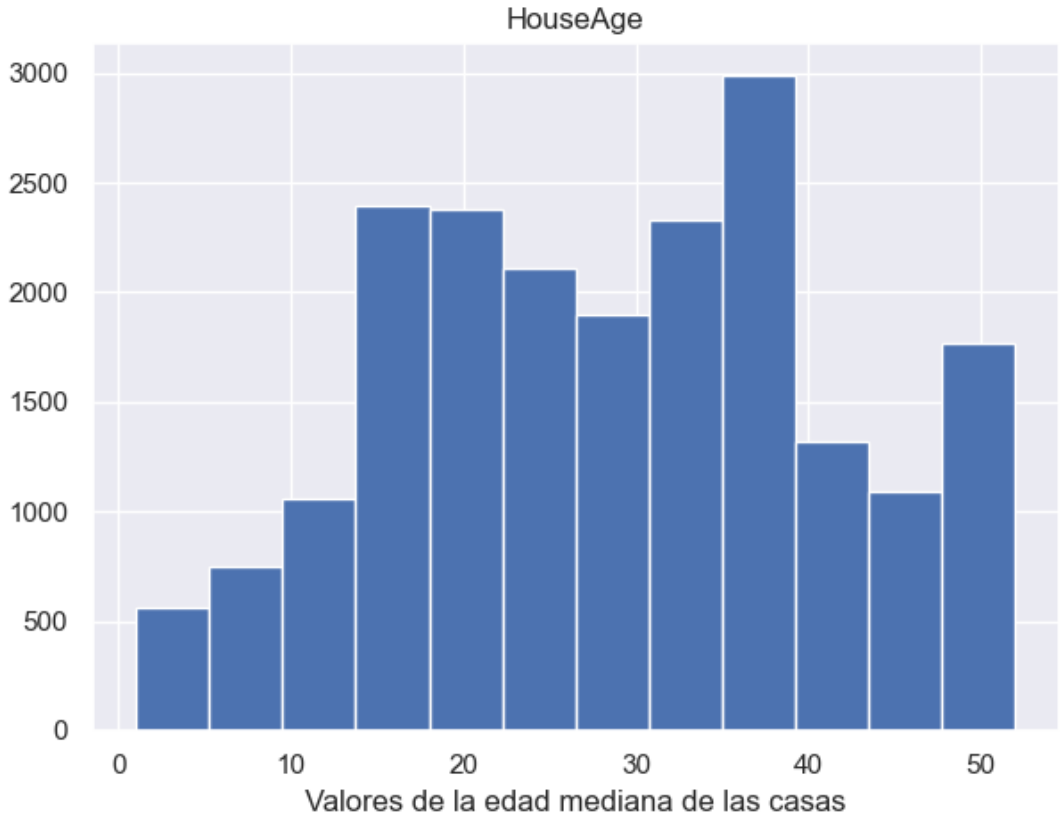
2. Graficar los histogramas de los diferentes atributos y el target. ¿Qué tipo de forma de histograma se observa? ¿Se observa alguna forma de campana que nos indique que los datos pueden provenir de una distribución Gaussiana, sin entrar en pruebas de hipótesis?

Histograma del ingreso medio en el bloque:



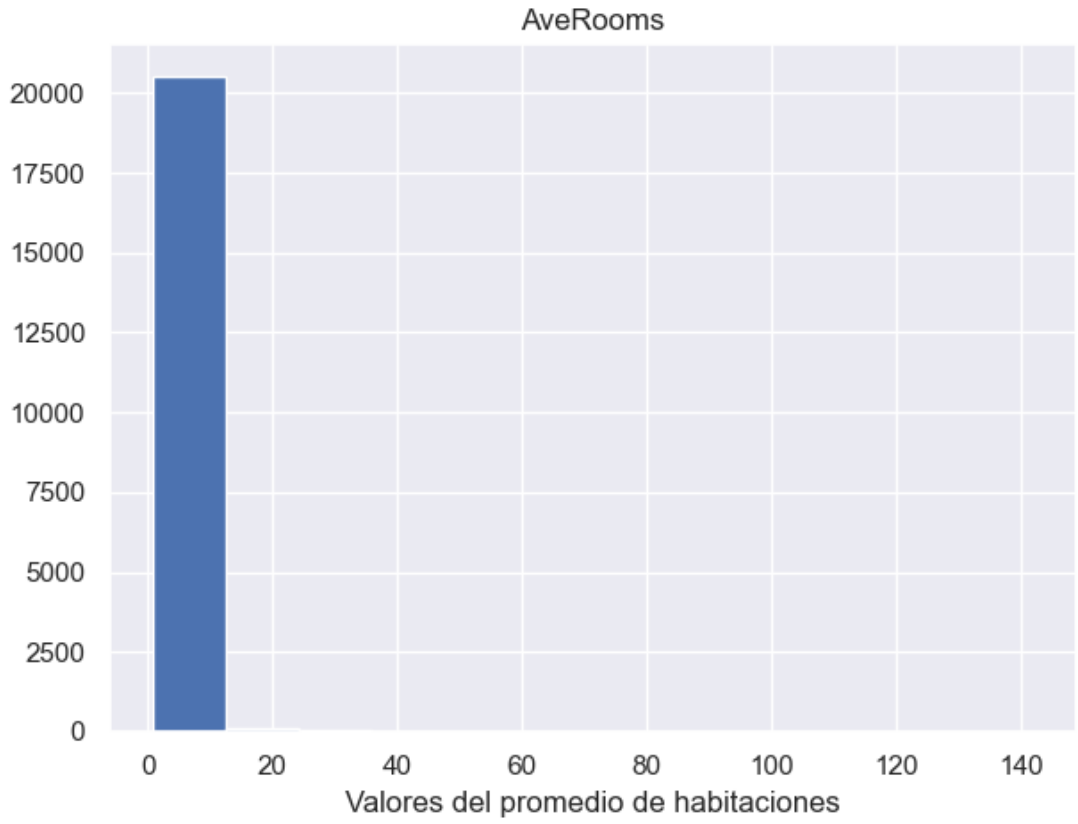
Para el caso del atributo de ingreso medio en el bloque, el histograma que se observa en efecto tiene una forma similar a la de una campana, lo cual nos indica que los datos pueden provenir de una distribución Gaussiana.

Histograma de la edad mediana de las casas del bloque:



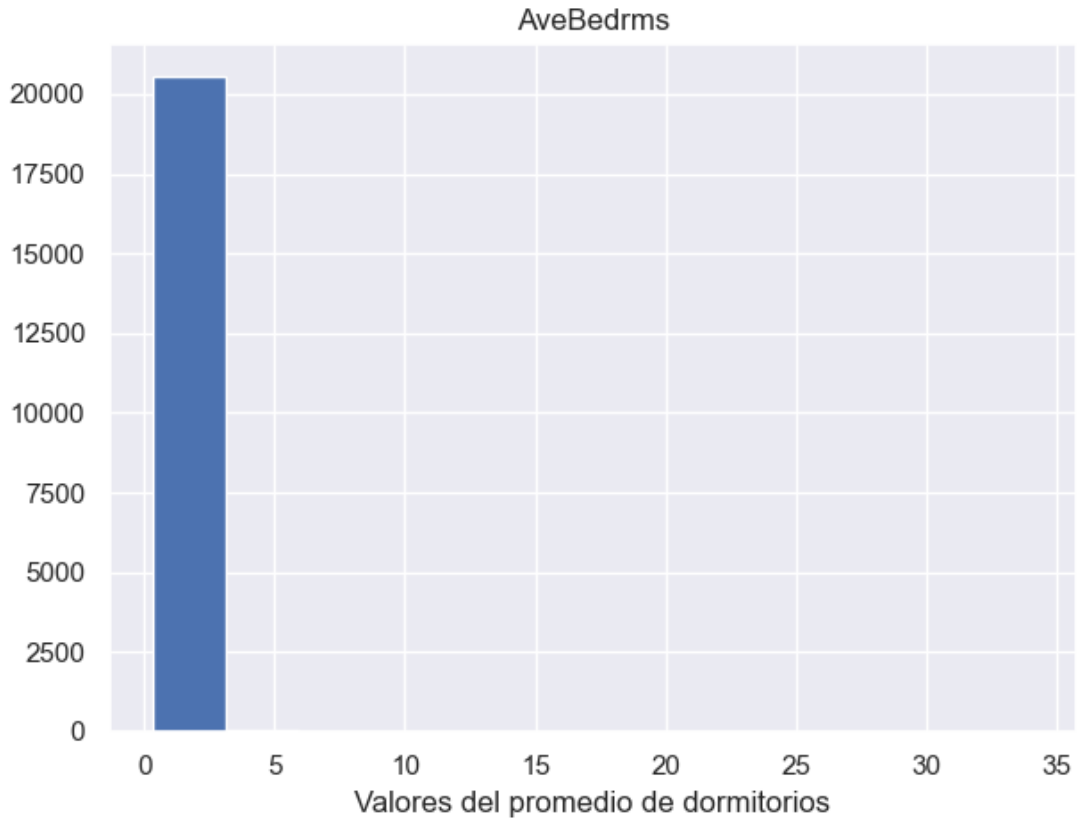
Con respecto al atributo de la edad mediana de las casas del bloque, el histograma presenta valores elevados de ocurrencias para edades de casas comprendidas entre los 15 a 25 años y también para las edades entre 30 y 40 años. Esto hace que el histograma tenga una forma de doble campana.

Histograma del promedio de habitaciones por hogar:



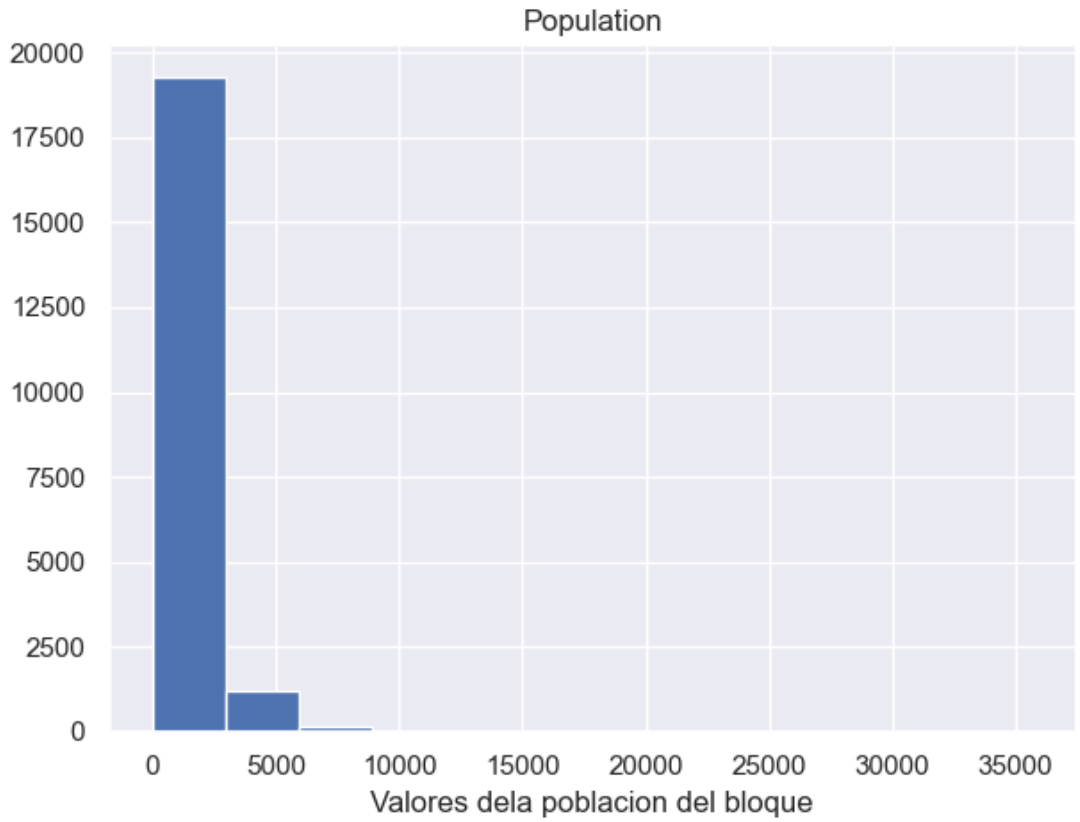
Con relación a los valores promedio del número de habitaciones por hogar, encontramos que la totalidad de estos valores se encuentran concentrados entre 1 a 12 habitaciones aproximadamente. Por lo que podemos ver que se trata de una distribución uniforme.

Histograma del promedio de dormitorios por hogar:



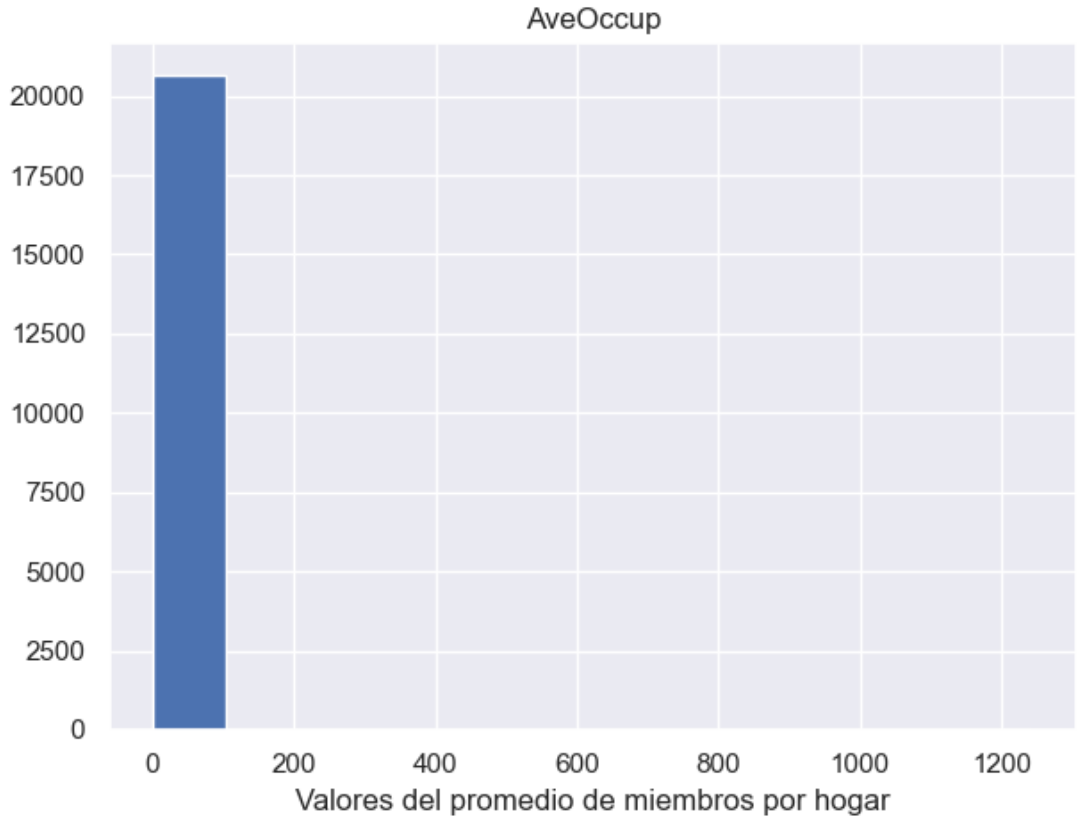
De forma similar al caso anterior, el histograma de promedio de dormitorios por hogar muestra que todos los valores se encuentran concentrados entre 1 a 3 dormitorios aproximadamente. Igual al caso anterior, se trata de una distribución uniforme.

Histograma de la población:



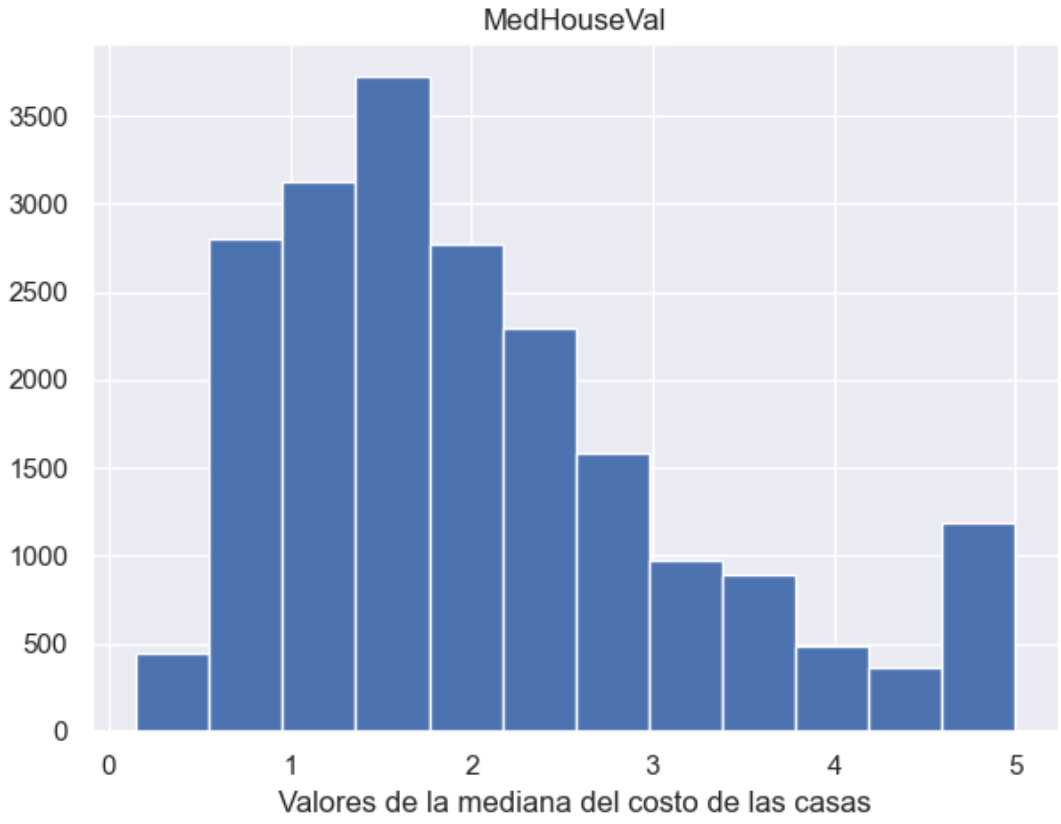
El histograma de la población, a pesar de tener la mayor cantidad de valores distribuidos en un rango mas acotado, también presenta unos cuantos valores cerca del rango de concentración mas alto. Por estos valores, la distribución también muestra una forma similar a la de una campana, por lo que se puede intuir que los datos también provienen de una distribución Gaussiana.

Histograma del promedio de miembros por hogar:



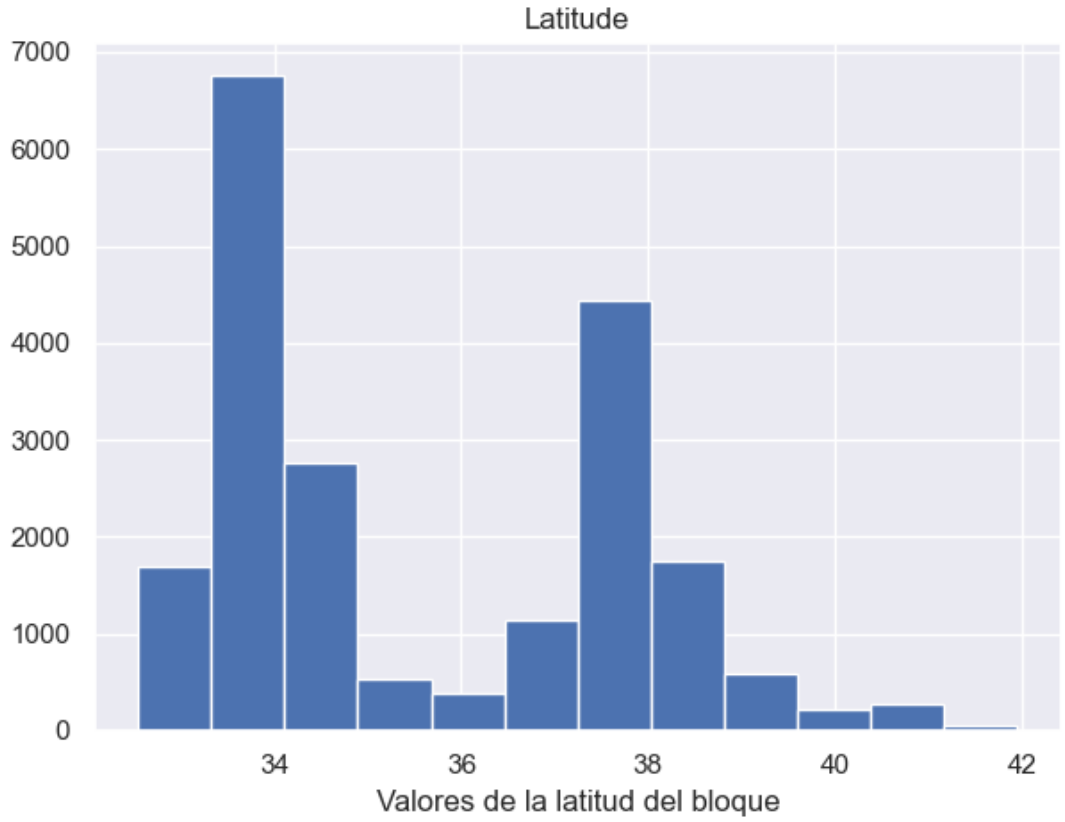
El histograma del número de miembros por hogar, muestra que la totalidad de los valores se encuentran concentrados entre los 0 y 100 miembros aproximadamente.

Histograma de la mediana del costo de casas en el bloque:



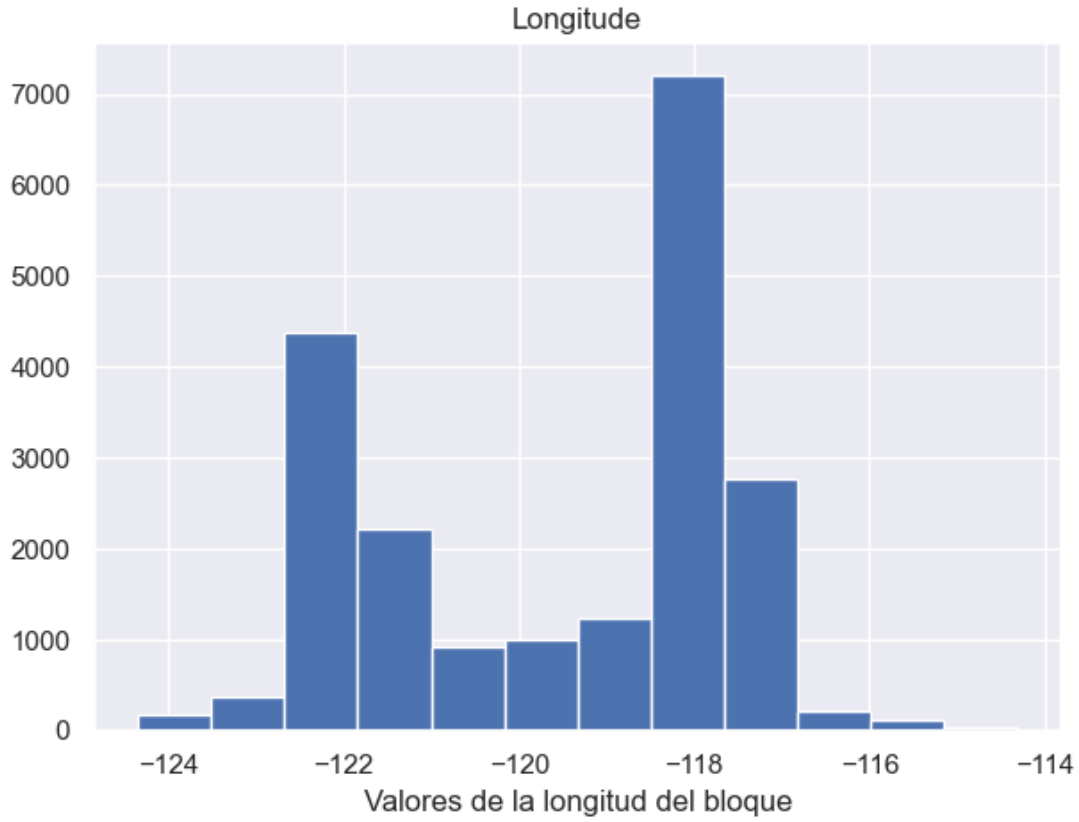
Para el caso de la mediana del costo de las casas en el bloque, que corresponde al target, se puede claramente apreciar una forma de campana en la distribución de los valores, indicando también que estos valores provienen de una distribución Gaussiana.

Histograma de la Latitud del bloque:



Para el caso de la latitud del bloque, se observa una forma de campana doble, lo cual también nos da indicaciones de que la latitud del bloque sigue una distribución Gaussiana.

Histograma de la Longitud del bloque:



Similar al caso anterior. Se observa que para la longitud del bloque también se tiene una forma de doble campana, indicando una distribución de tipo Gaussiana.

3. Calcular la regresión lineal usando todos los atributos. Con el set de entrenamiento, calcular la varianza total del modelo y la que es explicada con el modelo. ¿El modelo está capturando el comportamiento del target? Expanda su respuesta.

Calculando la regresión lineal, obtenemos:

- El valor de intersección de la recta es: 2.0692396089424165

- Los valores de los coeficientes de la recta son:

[8.49221760e-01 1.22119309e-01 -2.99558449e-01 3.48409673e-01
 -8.84488134e-04 -4.16980388e-02 -8.93855649e-01 -8.68616688e-01]

Obtenemos la varianza del modelo:

- La varianza del modelo es: 0.5235750603302349
- La varianza respecto del target es: 1.3396959774719193

Predicción del modelo:

La predicción del modelo es:

```
0       0.726049
1       1.767434
2       2.710922
3       2.835147
4       2.606958
         ...
6187    2.219941
6188    0.910516
6189    2.074655
6190    1.573714
6191    1.827441
```

Observando los valores de la predicción del modelo, se pueden identificar valores que concuerdan con lo observado en los valores del target, por lo que podemos concluir que el modelo en efecto refleja el comportamiento del target. Sin embargo, cabe notar que de la predicción del modelo expresada por el programa, solo se puede apreciar una parte de los valores.

A pesar de esta limitación, se observa que los valores varian dentro del rango observado en el comportamiento del target.

4. Calcular las métricas de MSE, MAE y R^2 del set de evaluación.

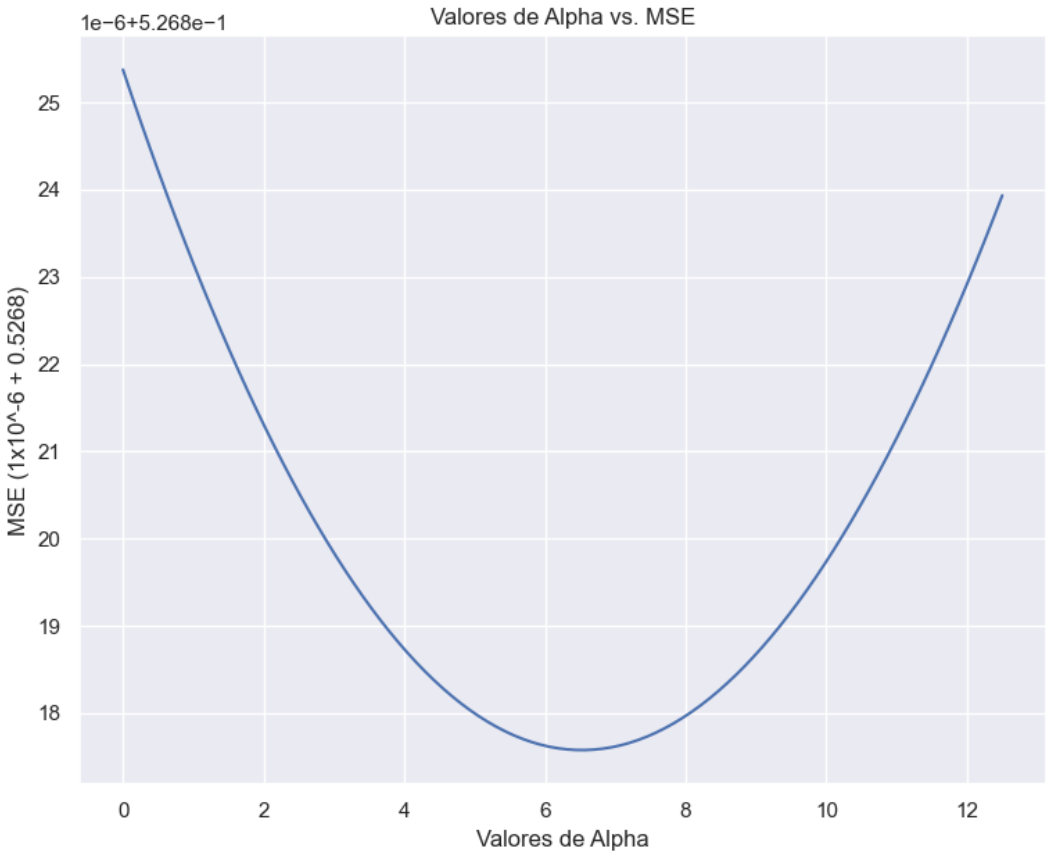
Obtenemos los siguientes valores:

- El error cuadrático medio (MSE) es: 0.5305677824766752
- El error absoluto medio (MAE) es: 0.5272474538305952
- El coeficiente de Pearson (r^2) es: 0.5957702326061665

5. Crear una regresión de Ridge. Usando una validación cruzada de 5-folds y usando como métrica el MSE, calcular el mejor valor de alpha, buscando entre [0, 12.5]. Graficar el valor de MSE versus alpha.

Aplicando la regresión con el modelo Ridge y los valores de Alpha entre 0 y 12.5, obtengo valores de MSE muy parecidos entre si, todos al rededor de 0.5268, donde solo los valores de 5to y 6to decimal presentan variación.

La gráfica obtenida es la siguiente:



Donde:

- El mejor valor de Alpha es: 6.565656565656566

- La media del MSE en 5-fold CV para la regresión Ridge con: alpha = 6.565656565656566 es: 0.5268175765319489

6. Comparar, entre la regresión lineal y la mejor regresión de Ridge, los resultados obtenidos en el set de evaluación. ¿Cuál da mejores resultados (usando MSE y MAE)? Conjeturar por qué el mejor modelo mejora. ¿Qué error puede haberse reducido?

Para la regresión Ridge, tomamos como referencia el mejor valor de Alpha (alpha = 6.5656) para obtener el mejor valor del error cuadrático medio (MSE), es decir MSE = 0.5268175765319489

Haciendo la comparación entre el valor de MSE obtenido con la regresión lineal (MSE= 0.5305677824766752) y el obtenido con el modelo de Ridge, se observa que el error cuadrático medio es menor utilizando la regresión de Ridge, es decir que hubo una mejora usándola regresión Ridge en comparación con la regresión lineal:

MSE(Ridge) = 0.5268175765319489
MSE(Regresión lineal)= 0.5305677824766752

Se puede conjeturar que el modelo que emplea la regresión de Ridge es mejor debido a la introducción del termino de penalización por encogimiento. El parámetro de ajuste “Alpha” con el que se obtiene el mejor valor del error cuadrático medio, corresponde a un valor que permite llegar a un equilibrio entre la varianza y el sesgo. Por este motivo es necesario probar entre un rango de valores de Alpha para encontrar el que permite este equilibrio, este valor también proporcionara el menor valor del error cuadrático medio.

NOTA: El repositorio donde se encuentra el código que generó los resultados descritos mas arriba se puede acceder en el siguiente link:

https://github.com/eduardo-echeverria/entregas-ai/blob/main/regresion_lineal/regresion_lineal_ridge.py