Instituto Nacional de Formación Técnica Profesional INFOTEP HVG

Análisis Univariado y Bivariado

Robert López, Dayanna Montes, Santiago Orozco

Resumen

El siguiente informe describe los análisis (univariado y bivariado) para cada uno de los datos analizados gráficamente. Para esta documentación usaremos 7 variables, cada una será gráficada en 3 casos como lo son: histogramas, medidas de centralidad y dispersión. Bajo este modo, mostraremos estadísticamente los datos y se redactará la interpretación que nos muestran las gráficas.

1 Introducción

Para este trabajo escrito haremos uso de la estadística, una rama de las matemáticas que nos ayudará a tener resultados aproximados o exactos sobre ciertos datos recolectados por medio de un censo, hecho en la ciudad de California. Aquí observamos los datos recogidos de manera grafica en cada una de ellas incluyendo la antigüedad de cada vivienda y lo que contienen cada una de la misma como por ejemplo: housing_median_age, total_rooms, total_bedrooms, population, households, median_income, median_house_value. Mediante estos datos vamos a construir las graficas que nos mostrara los items declarados en los objetivos de analisis.

2 Objetivos

2.1 Objetivo general

Crear un conjunto de graficas mediante los análisis univariado y bivariado haciendo uso de 7 variables, con el fin de comprender sus resultados e interpretaciones de manera correcta.

2.2 Objetivos específicos

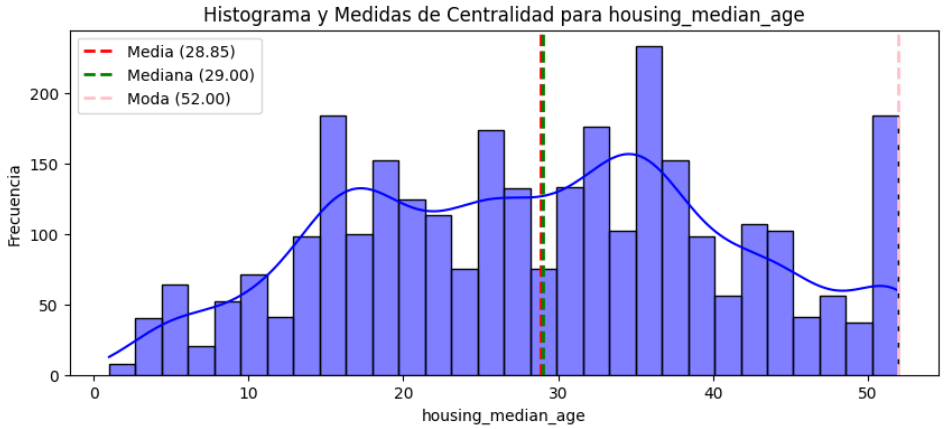
- Comparar datos precisos sobre la antigüedad, valor y otros aspectos a traves del histograma haciendo uso de las graficas.
- Mostrar mediante la medida de centralidad los valores sobre las variables mas comunes o menos comunes segun los metodos de la estadísticas.

- Visualizar mediante las graficas de dispersión los lugares donde se agrupan las variables segun su clase y donde son menos notables.

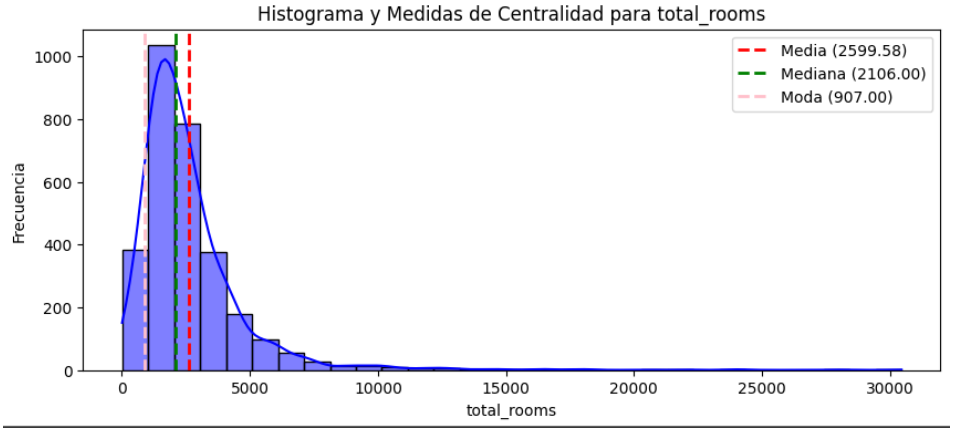
3 Descripción de los Análisis (Univariado)

A continuación vamos analizar las gráficas generadas en Jupyter Notebook. Para este primer análisis, el cual es el análisis univariado nos ayudaremos de histogramas, medidas de centralidad y medidas de dispersión para hacer las respectivas gráficas y su interpretación.

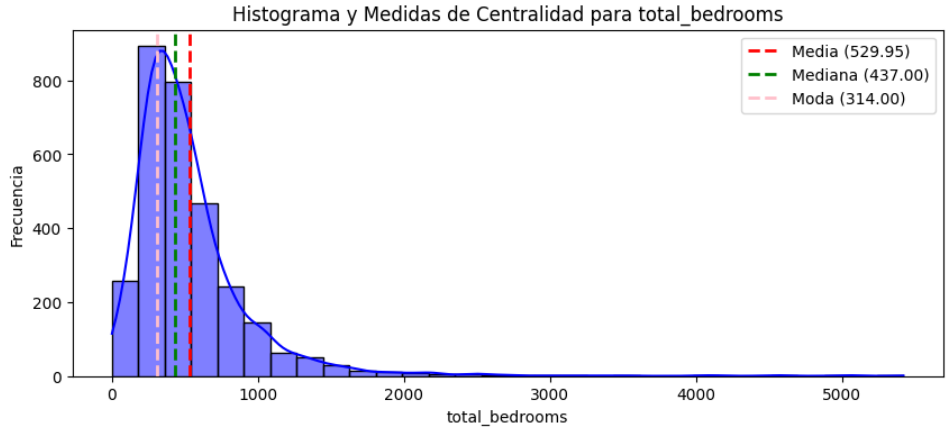
3.1 Gráficas de resultados



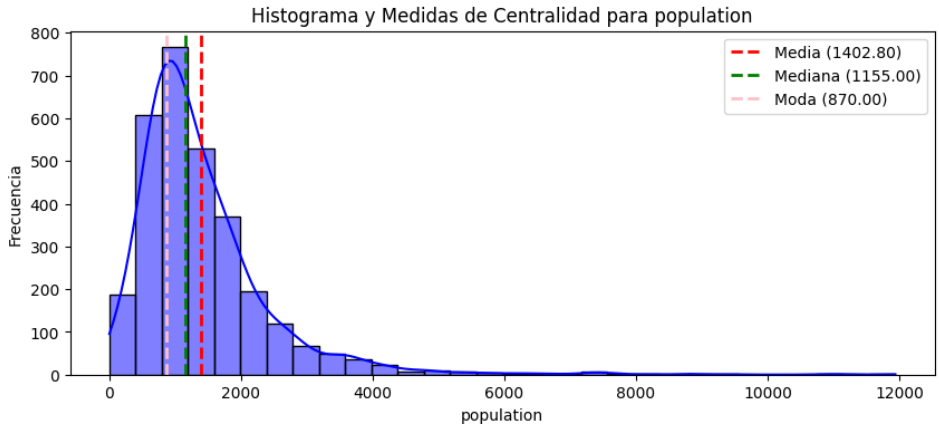
Analizamos en la primera gráfica la primera variable housing_media_age que equivale a la edad promedio de las casas en california la imagen muestra que la media se encuentra en el eje X (28.85) lo que quiere decir que el promedio de las casas indica esa antigüedad por lo tanto podemos decir que la menor edad de las casas de esta ciudad es de 28.85. También podemos ver que la mediana se encuentra en (29.00) cerca de la media. Y la moda en (52.00) lo que quiere decir que la ciudad contiene un numero mayor de casas antiguas.



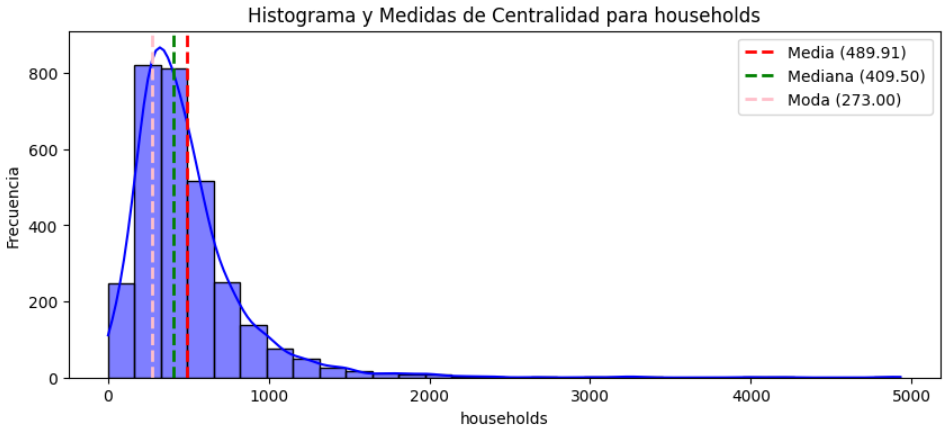
En la gráfica anterior representamos los valores de los dormitorios por vivienda, la imagen nos muestra estadísticamente que tenemos un número promedio de total_rooms por cada casa y también observamos que la mediana es similar a la media por otro lado observamos que la moda es más frecuente sin llegar a afectar la medida de la mediana y la media.



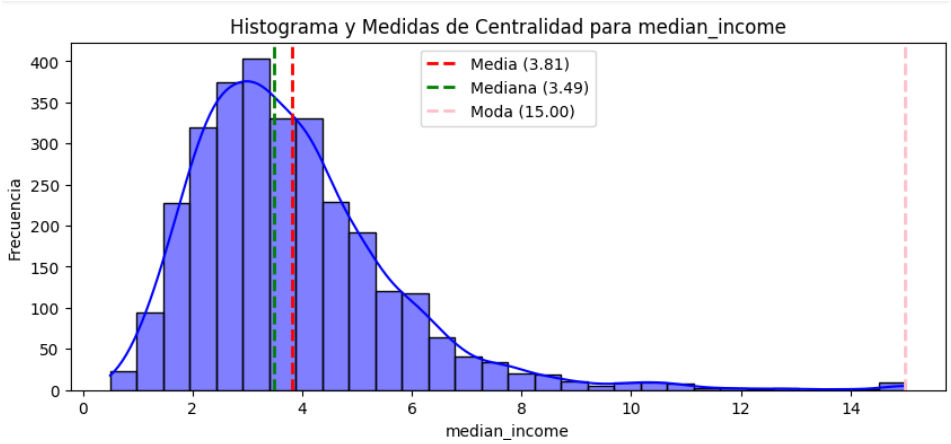
Podemos observar en esta gráfica el número total de habitaciones por cada vivienda se visualiza que la media se encuentra mucho más alto (529.95) que la mediana (437.00) un poco más centralizado. Por otro lado tenemos la moda en (907.00). Lo que quiere decir estadísticamente que tenemos más casas con un número de habitaciones mayor a lo que seria mas frecuente según la moda.



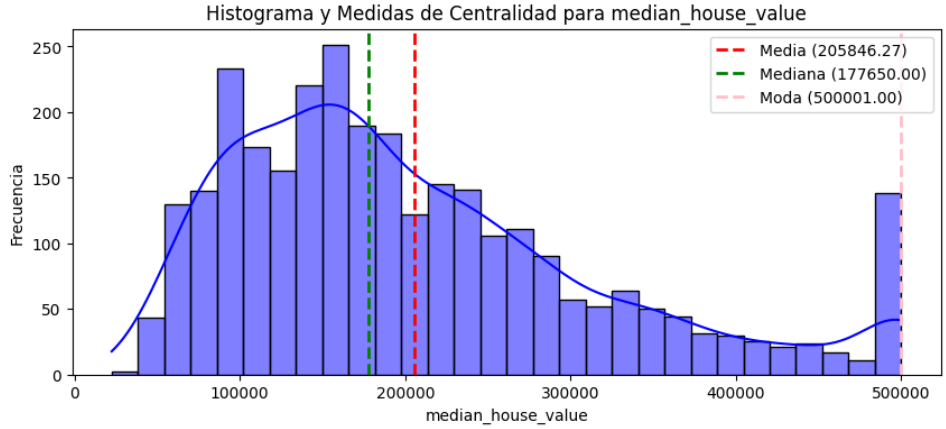
Para la gráfica population tenemos los siguientes valores media (1402.80), mediana (1155.00) y la moda (870.00) lo que nos hace pensar que en los dos análisis anteriores (total_rooms y total_bedrooms) concuerda con el número de population en la vivienda es decir, que a mayor número de población mayor cantidad de habitaciones y dormitorios, tal cual como se muestra en la imagen.



En la gráfica anterior tenemos representada la variable households donde se define estadísticamente la cantidad de hogares en la ciudad, tenemos la media con (489.91), la mediana con (409.50) y la moda con (273.00) estos valores nos permite comparar la gráfica con la gráfica de la variable population, dejándonos saber que la media de hogares es perteneciente a familias numerosas, aunque la moda nos indica que hay un número de hogares con números de personas más bajo, pero al mismo tiempo más frecuente.

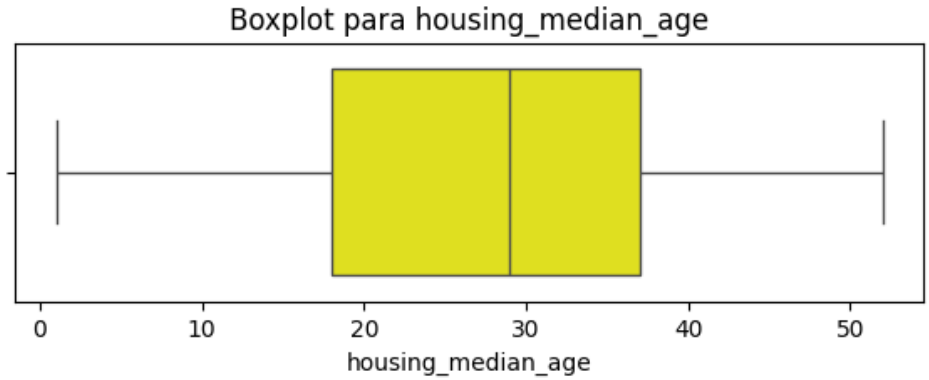


En la gráfica anterior analizamos el median_income lo que también podemos decir entrada económica, así pues tenemos que el median_income tiene media (3.81), mediana (3.49) y la moda (15.00). La gráfica nos representa estadísticamente que la mitad de los datos recolectados tienen una entrada económica muy cercana al promedio, pero muy lejos de lo frecuente así como lo muestra la moda.

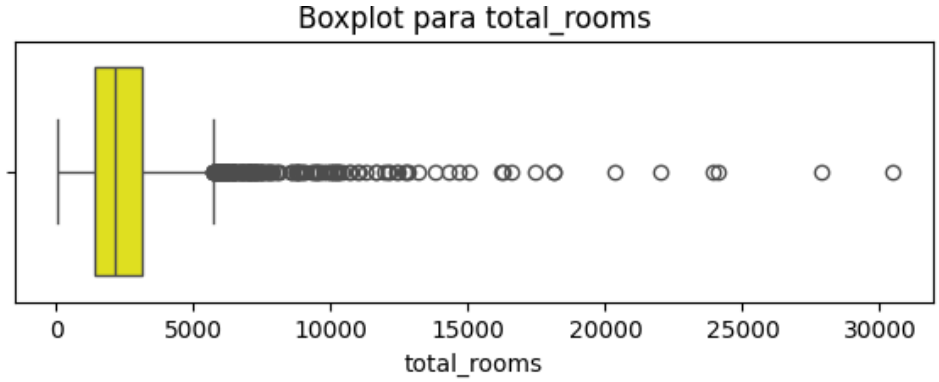


Para la anterior gráfica podemos decir lo siguiente, observamos la siguiente estadística tenemos la moda con (500001.00) lo que nos indica que es muy frecuente que hayan casas valorizadas a lo mencionado anteriormente, pero la mediana nos indica que la mitad de los datos recolectados sobre la vivienda tienen un costo menor que a la media, así como se visualiza en la imagen.

3.2 Boxplot



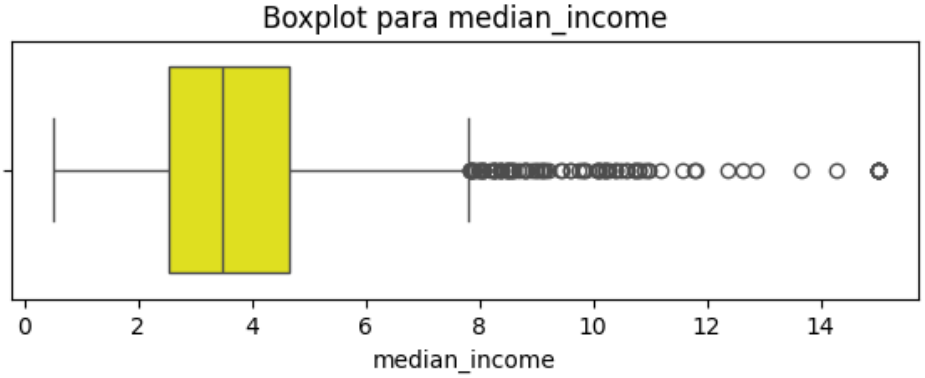
En esta gráfica de boxplot se muestra una amplia distribucion de la edad mediana en las viviendas, habiendo una concentracion importante entre 18 y 34 anos, y con esto unos valores minimos y maximos que van de alrededor de 2 anos, hasta cercanas a 50.



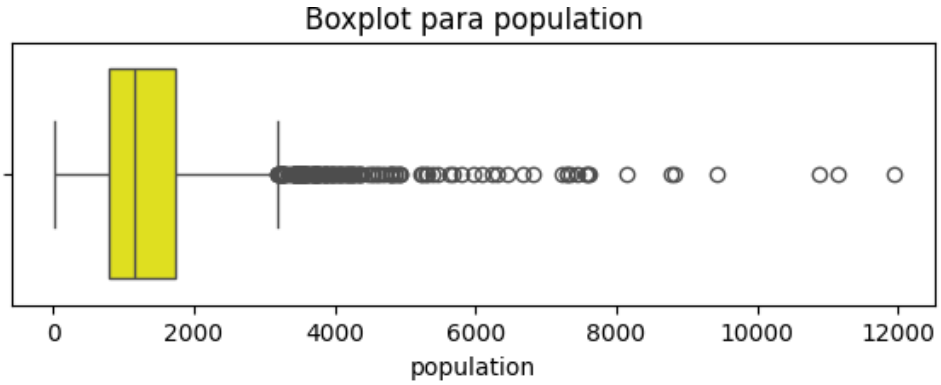
En la grafica con variable total_rooms se muestra que los datos no estan distribuidos de manera uniforme en ambos lados, lo que conlleva a una gran cantidad de datos concentrados en rangos bajos cercanos a 0 y otro conjunto de valores muy altos que llegan a 30,000, lo que se visualizan como valores atípicos, haciendo extender notablemente la escala en la gráfica.

En la gráfica con variable total_bedrooms se muestra que la mayoria de las propiedades tienen un total de recamaras relativamente bajo de 1000, pero tambien hay un subconjunto de propiedades que llegan a valores elevados de 5000 recamaras. Lo que explica por que la media de distribucion podria ser mayor que la mediana, ya que unos pocos valores grandes jalan la distribución hacia la derecha. Lo que se ve reflejado en que alla presencia de propiedades con muchas mas recamaras que el promedio

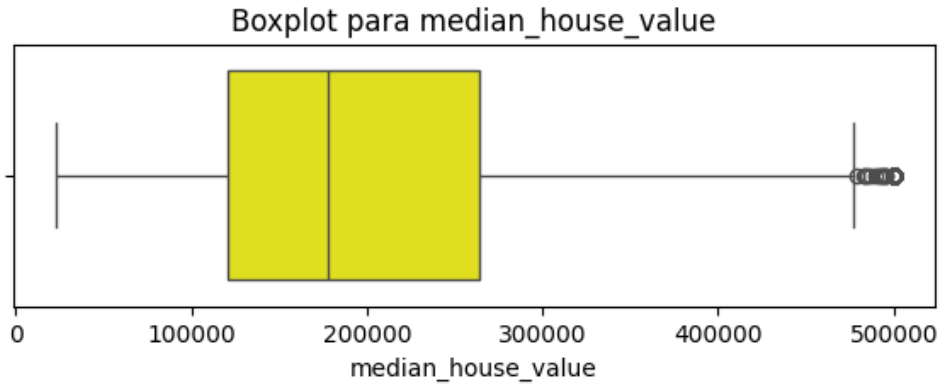
Para la gráfica anterior tenemos el caso de dispersión con la variable households esta nos representa el mayor número de hogares agrupados por zona. Podemos decir que el eje X nos representa la cantidad de hogares donde se realizo el censo y los puntos con dispersión representa las zonas donde se recolectaron los datos, es decir que las zonas mas lejos de la cantidad agrupada es un valor atípico.



En la gráfica con variable median_income se interpreta que la mayor parte de los ingresos medianos se concentra en torno a la franja de 2 a 5, con una mediana cerca de 3-4, pero hay subconjuntos con zonas de ingresos medianos muy elevados lo que provoca que los ingresos no esten distribuidos de manera uniforme y con una media por encima de la mediana.

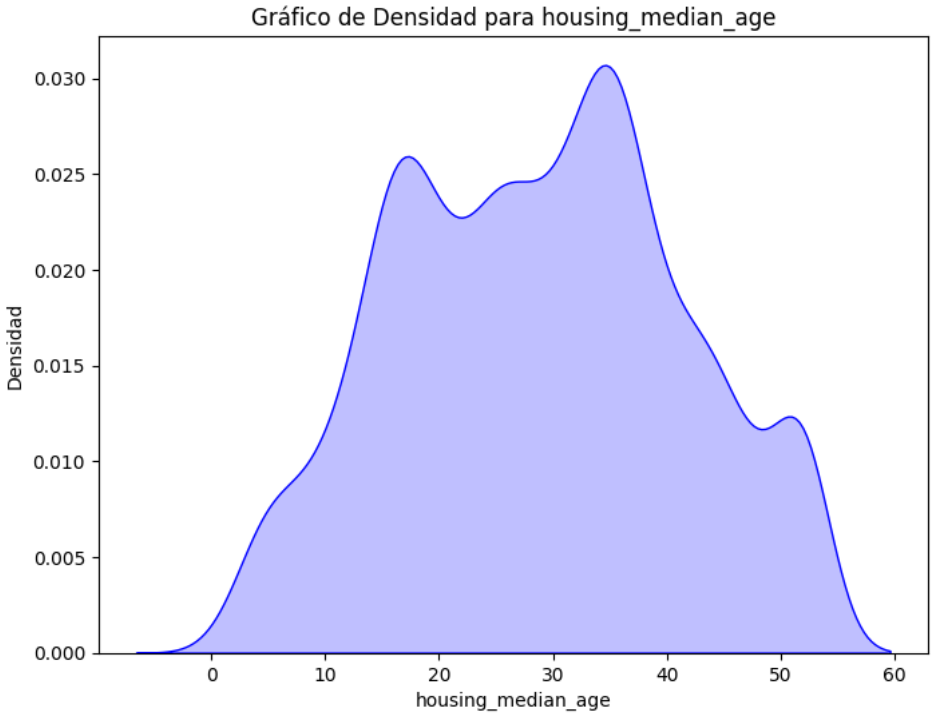


En la gráfica con variable population se observa que hay un rango muy amplio con valores minimos cercanos a cero y maximos que superan los 10 mil habitantes lo que nos indica que hay zonas con muy poca poblacion y otras con mayor poblacion de habitantes. También se observa que la caja representa el 50% central de datos, que esta aproximado entre los 800 y 2000 habitantes con una mediana que ronda los 1500 habitantes, lo que da entender que la mitad de las zonas tienen una poblacion menor a esa cifra y la otra mitad, mayor.

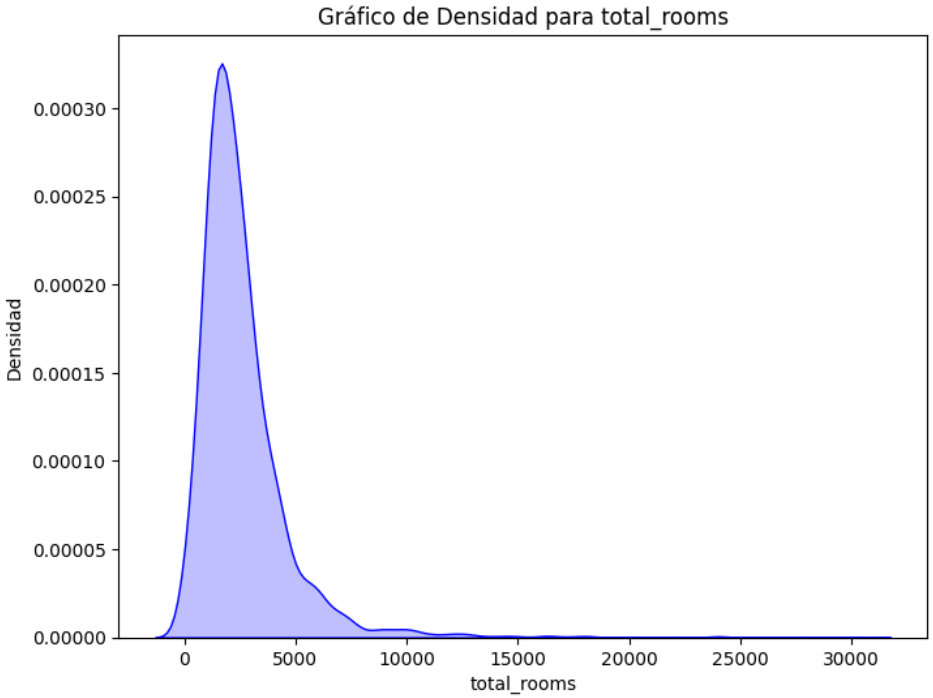


En la gráfica con variable median_house_value se ve que la mayor parte los valores de una vivienda se concentran en un rango aproximado de 120.000 a 250.000 dolares, con una mediana cercana entre 180.000 y 200.000 dólares. También hay un grupo de zonas con valores medianos altos superiores a 400.000 dolares, hasta llegar a un limite de 500.000 dolares, en conclusión el patron sugiere que hay una distribucion en cola larga hacia valores elevados, lo que puede arrastrar a la media por encima de la mediana y hace visible la asimetría en la gráfica.

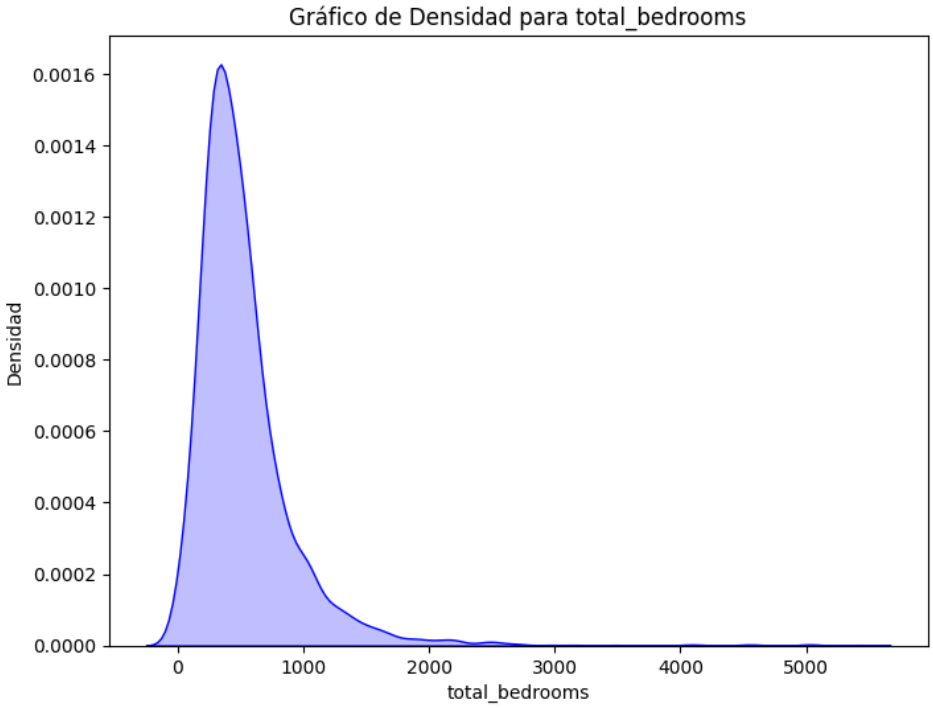
3.3 Gráfico de densidad



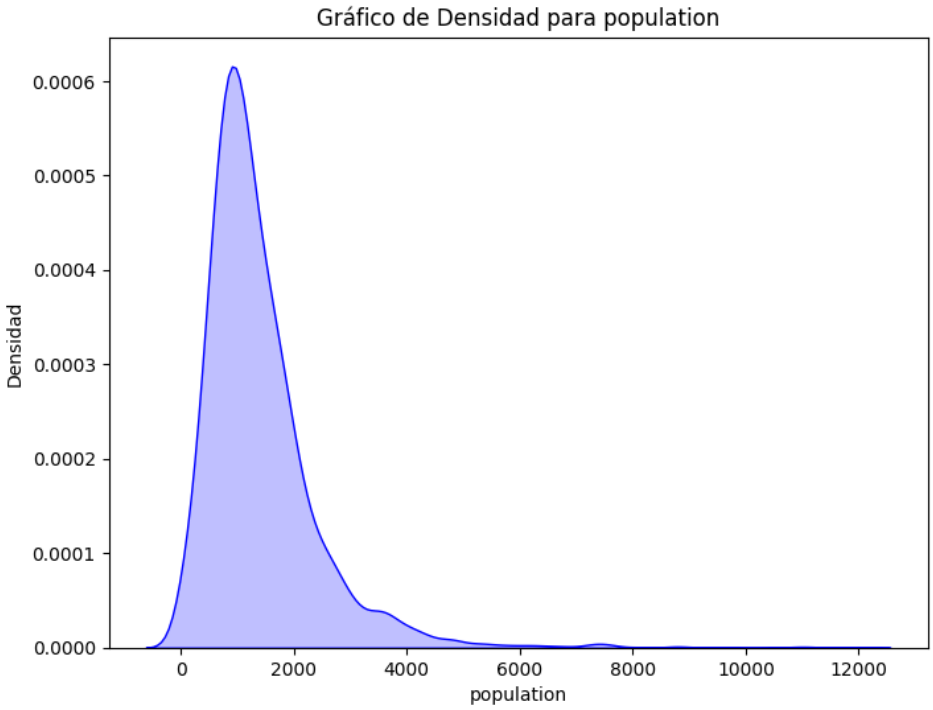
Podemos ver que la gráfica tiene unos puntos de ascenso y otros donde cae, por ejemplo, se puede ver un ascenso antes de llegar a los 20, vuelve y tiene un repunte alrededor de los 25, por último alrededor de los 50 vuelve y asciende. Podemos observar que la mayor densidad va entre 20 y 40 años, es decir que la mayoría de viviendas tienen una edad mediana dentro de ese rango.



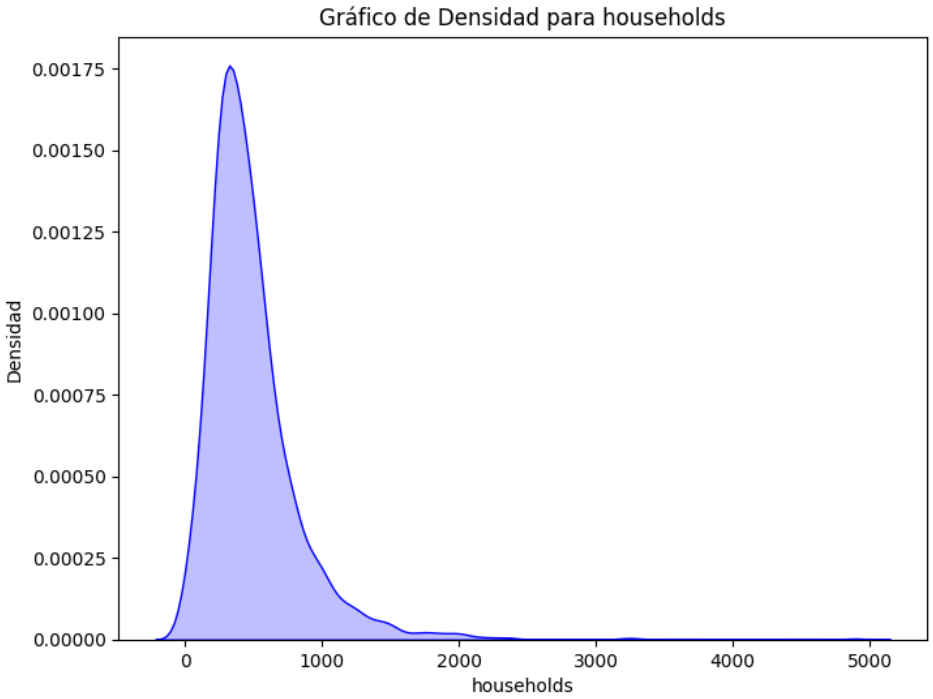
Se puede observar que la anterior gráfica tiene un ascenso entre 0 y 3000, de ahí desciende y sigue de largo hasta los 30,0000 y más. La mayoría de los datos están ubicados en valores relativamente bajos, por otra parte los valores mas altos se extienden en una cola hasta muchísimas más habitaciones.



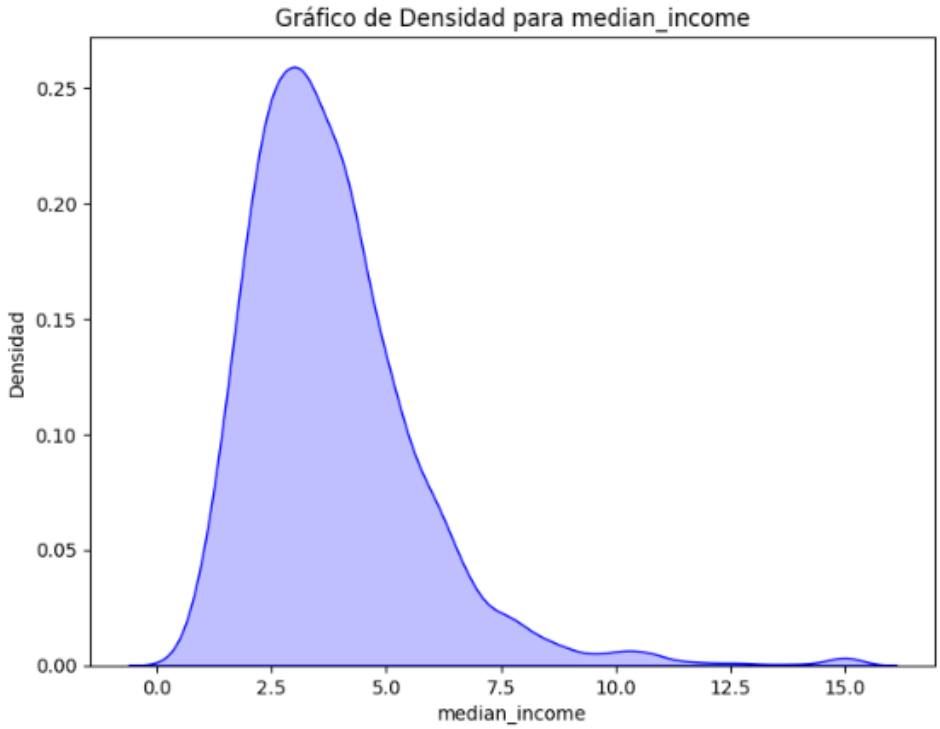
La gráfica de densidad de la variable total_bedrooms se observa que tiene una concentracion principal de valores bajos, pero con una cola extendida hacia la derecha lo que provoca un alto rango que incrementa notablemente la varianza y desviación estándar lo que marca una asimetría. Lo que sugiere que la media sera mayor a la mediana y que las medidas de dispersión sensibles a valores maximos como la varianza y la desviación estándar seran bastante altas.



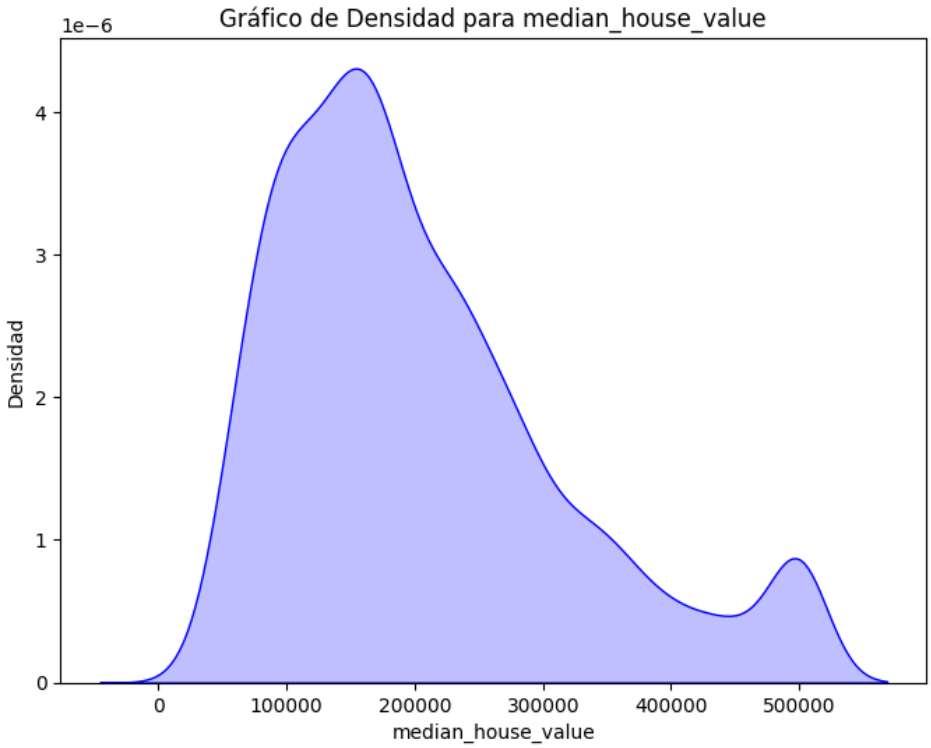
En la anterior gráfica podemos observar que tiene un repunte entre 0 y 1500, luego desciende con una cola hasta los 12000 o más, lo que quiere decir que la presencia de registro con población más alta. La gráfica de population podemos ver que tiene una mayor densidad en valores más bajos y menos densidad en los valores altos.



En la anterior gráfica se puede ver un pico que va de 0 a 500 aproximadamente, luego desciende con una cola hasta los 5,000. Podemos ver que la densidad más alta se encuentra donde están pocos hogares, pero un número reducido de zonas o registros presentan valores muy altos.



La gráfica de densidad de la variable median_income presenta una distribucion unimodal, con una fuerte concentracion en valores bajos a moderados y una cola derecha que se extiende a valores elevados, lo que provoca una distribucion asimétrica y con altas medidas de dispersión cuando se utiliza indicadores sensibles a valores extremos como la varianza y la desviación estándar debido a la cola larga hacia la derecha.

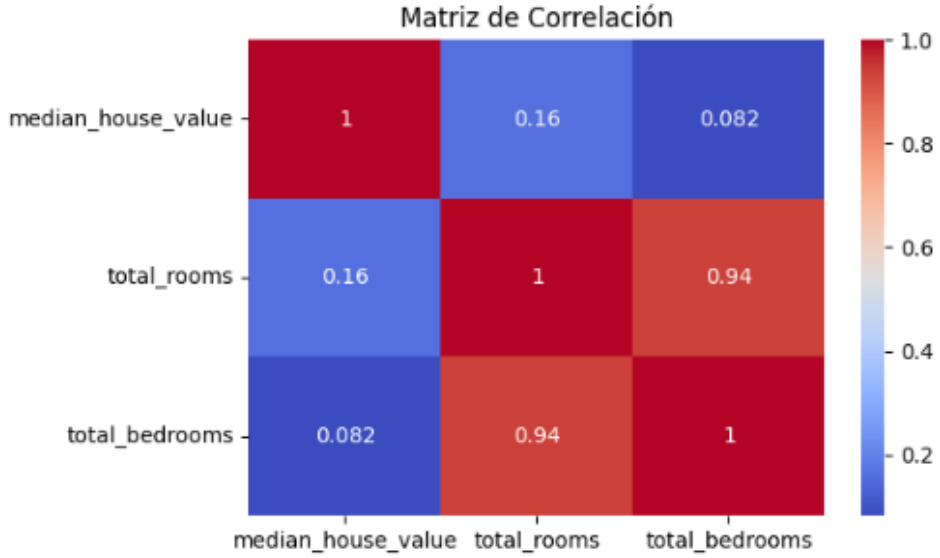


La gráfica de densidad de la variable median_house_value muestra que las propiedades que tienen valores medianos de vivienda en un rango moderado por debajo de 300.000 dolares, con un pico principal alrededor entre 100.000 y 200.000 dolares. No obstante hay un segundo grupo de observaciones que se concentra en valores altos entre 400.000 y 500.000, lo que crea una cola extendida y un posible segundo pico de 500.000. Lo que genera un mercado inmobiliario con gran variabilidad y la posible censura en el registro de precios, lo que hace que todas las viviendas con valores por encima de 500.000 podrian estar agrupadas en ese punto maximo.

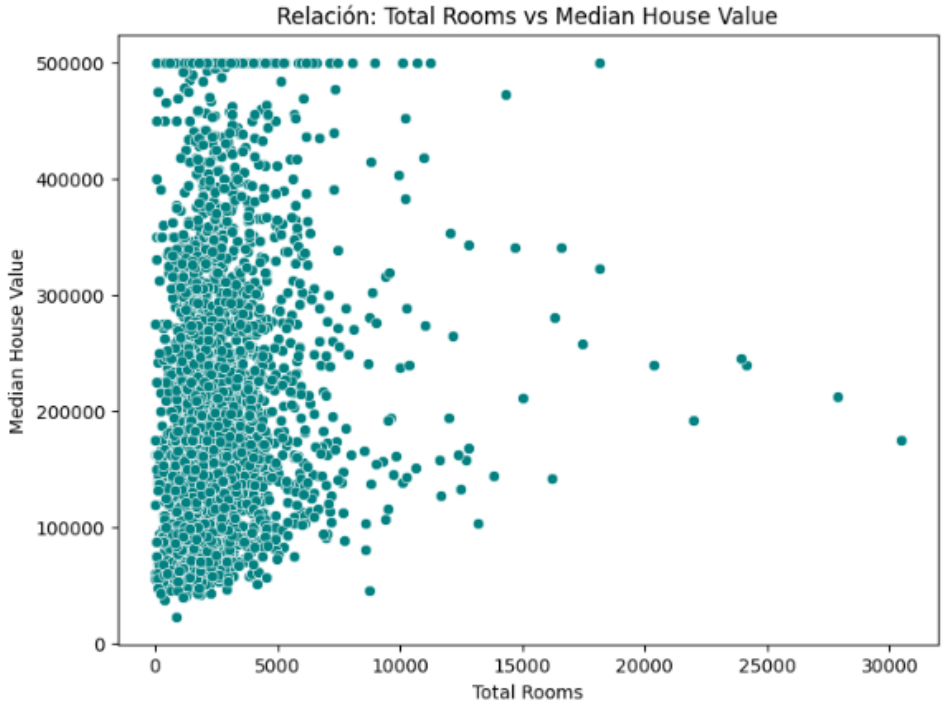
4 Descripción de los Análisis (Bivariado)

A continuación vamos analizar las gráficas generadas en Jupyter Notebook. Para este segundo análisis, el cual es el análisis bivariado o multivariado usaremos las siguientes variables housing_median_age, total_rooms, median_house_value para determinar posibles patrones de correlación.

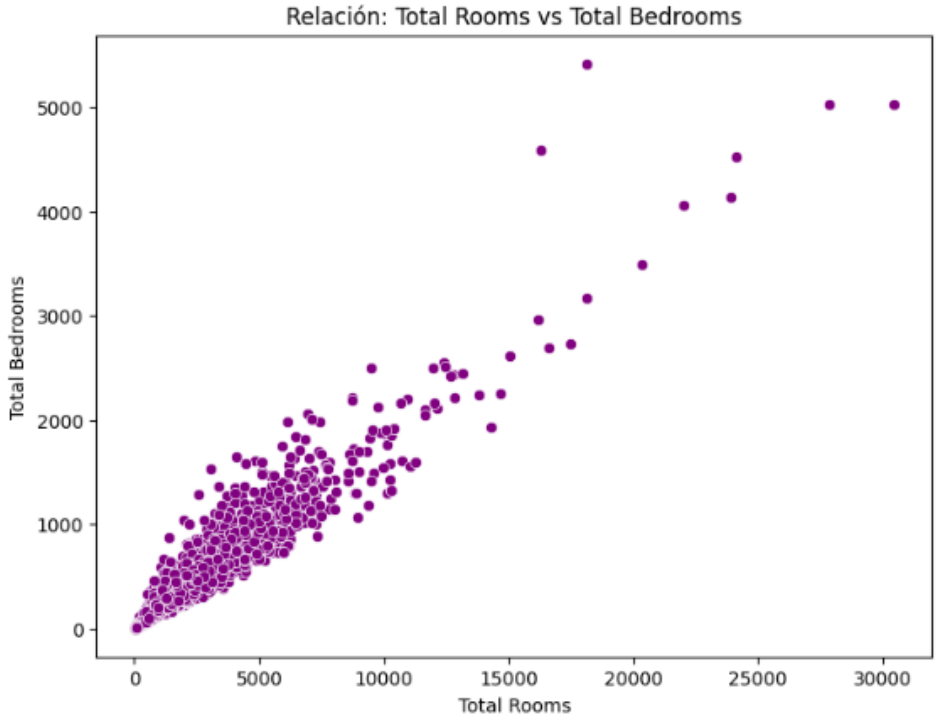
4.1 Gráficas de resultados



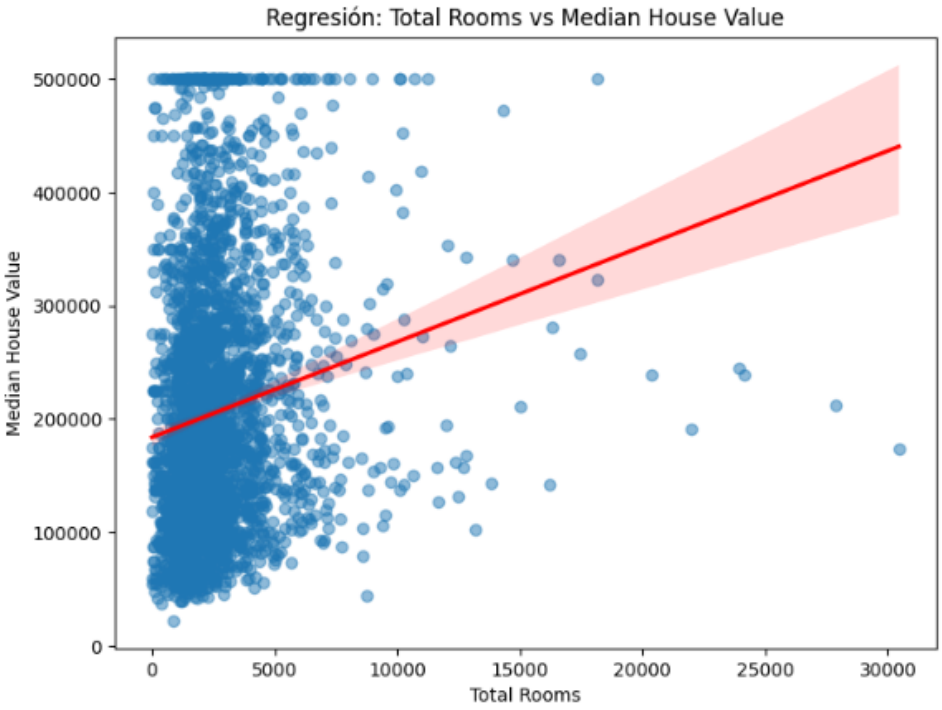
Podemos ver una fuerte correlación entre total_bedrooms y total_rooms con un puntaje 0.94, lo que quiere decir que se relacionan muy bien, pues entre más habitaciones haya, la probabilidad de que haya dormitorios es mayor. Se ve una débil correlación entre median_house_value y total_bedrooms con un puntaje de 0.082. Es decir que el valor de la casa no esta 100% relacionada con las habitaciones, depende de otros aspectos. Y hay una baja correlación entre median_house_value y total_rooms con un puntaje de 0.16, lo cual no se correlacionan en nada, ya que el valor de la casa depende de otros factores.



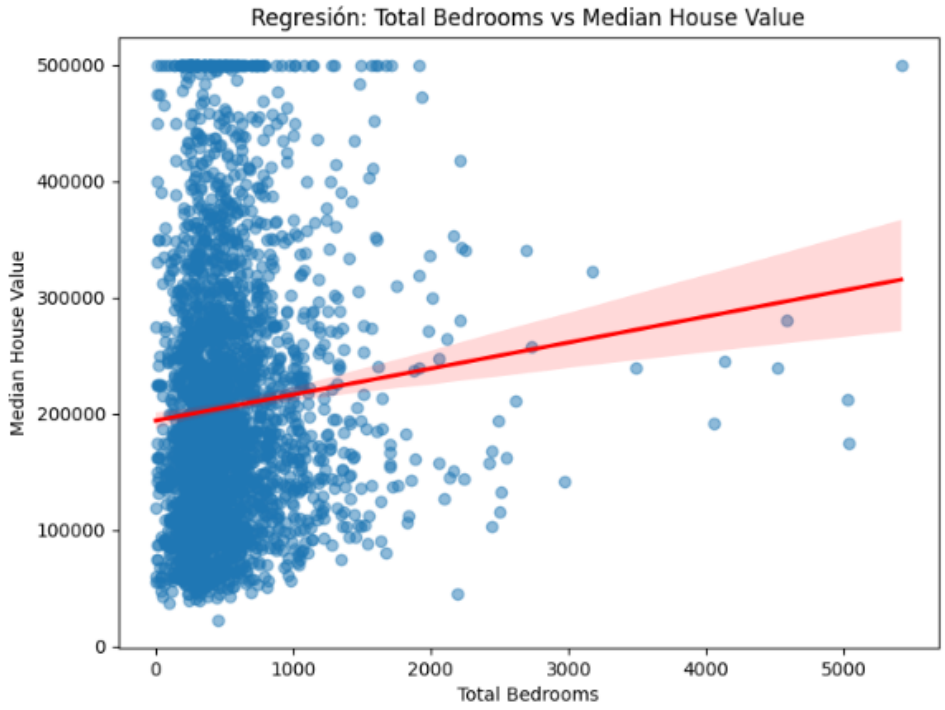
La gráfica de total rooms vs Median House Value se puede observar que hay una relacion positiva moderada, a mayor numero de habitaciones, mayor valor de la vivienda. Pero no seria determinante debido a la gran dispersión, por otro lado debido a la concentracion de muchos puntos en zonas de pocas habitaciones con valores de viviendas muy diversos, indica que el total_rooms no seria el unico predictor potente para explicar el valor mediano de la vivienda. Esto sugiere que si bien existe una tendencia a que las viviendas con mas habitaciones sean mas costosos, no seria un factor determinante por si solo y estaria fuertemente condicionado a otros elementos del contexto.



En la anterior gráfica podemos ver que hay un mayor de puntos en la parte inicial de los ejes, pues sabemos que si crece el número de habitaciones, igualmente crece el número de dormitorios. Lo que habíamos dicho anteriormente en la matriz de correlación. El eje X va hasta 30,000 dormitorios mientras que el eje Y va hasta 5,000 habitaciones. La mayoría de los puntos se concentran en la parte baja e intermedia, pero existen algunos valores muy elevados de zonas con un número muy alto de cuartos y dormitorios.



La gráfica de total Rooms vs Median House Value se puede observar una tendencia positiva en la linea de regresión que presenta una pendiente positiva, lo que sugiere que, a medida que aumentan las habitaciones totales en la zona, tiende a crecer el valor mediano de la vivienda, tambien se encontro que hay alta concentracion de datos en la parte baja por un gran acumulo de puntos en la zona de Total Rooms menor que 5.000, lo que abarca un rango muy amplio de valores de vivienda. Hay puntos donde se superan los 10.000 0 20.000 total_rooms, que se reparten en valores de vivienda desde 100.000 hasta 500.000. Surge un tope de 500.000 dolares, lo cual en muchos conjuntos de California corresponde aun limite de censura lo que genera una acumulacion de puntos en esa zona. En conclusión se aprecia una relacion positiva entre el numero total de habitaciones y valor mediano de la vivienda, lo que hace que la correlación no se muy fuerte, por lo tanto total_rooms puede aportar información parcial sobre le valor mediano de la vivienda, pero no resultaria preditorio por si solo.



La siguiente observación es para la gráfica de total_bedrooms vs median_house_value. Los datos mostrados en el eje X representan los números de habitaciones en una vivienda y el eje Y nos representan el valor promedio de la casa. También vemos una línea que atraviesan la agrupación de puntos. Lo que nos da a entender que el valor de la casa va subiendo a la medida que tiene más habitaciones. Pero este costo no solo se basa en el número de habitaciones, sino que también hay otros factores fundamentales que influyen en este valor. Por lo que hay varianza entre los precios, así como nos indica la dispersión de puntos.

5 Conclusión

Esta documentación se concluye diciendo que la información gráficada puede darnos una visualización más general de los datos propuestos en los casos, y estos a su vez nos indica cuales son los valores más variados o los más comunes en el lugar donde se tomó la muestra. Gracias al análisis univariado (técnica estadística que se utiliza para estudiar una sola variable de un conjunto de datos. Su objetivo es describir y resumir datos para descubrir patrones) y al analisis bivariado (método estadístico que estudia la relación entre dos variables. Se utiliza para obtener datos estadísticos sobre la influencia que tienen las variables entre sí). Se pudo establecer la relación entre variables, patrones comunes y obtener una mejor interpretación

con cada una de las variables graficadas.