Análisis en python sobre el precio de inmuebles en Washington

Autor: Cabrera Gabriel

AGENDA

- 01 | Contexto y Audiencia
- 02 | Hipótesis
- 03 | Metadata
- 04 | Análisis Exploratorio
- 05 | Insights y Recomendaciones
- 06 | Feature Selection

CONTEXTO Y **AUDIENCIA**

Contexto

En el contexto de Washington, Estados Unidos, el negocio de bienes raíces ha experimentado un crecimiento significativo en las últimas décadas. La región se ha convertido en un destino atractivo tanto para residentes locales como para inversores internacionales debido a su próspera economía, oportunidades laborales y calidad de vida. El mercado inmobiliario en Washington ofrece una amplia gama de propiedades, desde apartamentos urbanos hasta lujosas casas suburbanas y fincas rurales.

Considerando el crecimiento del mercado inmobiliario en Washington, existen diversas oportunidades de negocio que motivan a ser exploradas:

A)Agencia inmobiliaria especializada en propiedades de lujo

B)Servicios de consultoría para inversionistas internacionales

C)Desarrollo de propiedades sostenibles

D)Servicios de remodelación y diseño de interiores

Audiencia

El análisis de este dataset sería relevante para emprendedores y profesionales del sector inmobiliario interesados en aprovechar el crecimiento constante del mercado inmobiliario en la región.

Podrían surgir oportunidades de negocio, como agencias especializadas en propiedades de lujo, servicios de consultoría para inversores internacionales, desarrollo de propiedades sostenibles y servicios de remodelación y diseño de interiores.

PREGUNTAS DE **INTERÉS**

Se plantearán tres hipótesis:

1. La primera hipótesis tiene como objetivo determinar si la presencia de una pileta en una casa influye en su precio y en qué medida.
2. La segunda hipótesis es encontrar indicadores que permitan predecir el precio en función del código postal de la zona.
3. Por último, se investigará si el año de construcción de una casa tiene un impacto en su precio de venta.

Se espera que los resultados obtenidos a través de este estudio permitan mejorar la precisión de las predicciones de precios de casas y proporcionen información valiosa para los actores del mercado inmobiliario en Estados Unidos.

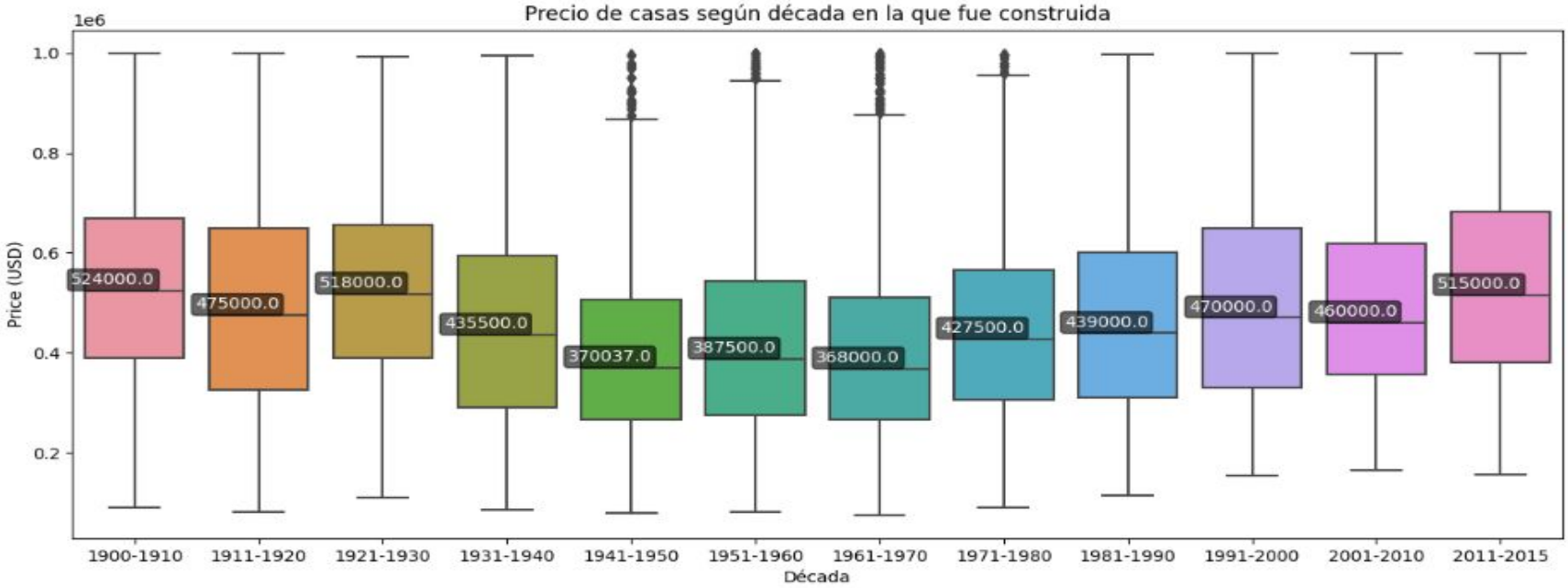
RESUMEN **METADATA**

+20 K
Casas Vendidas en 2 años

$999.999
Precio Máximo

$75.000
Precio Mínimo

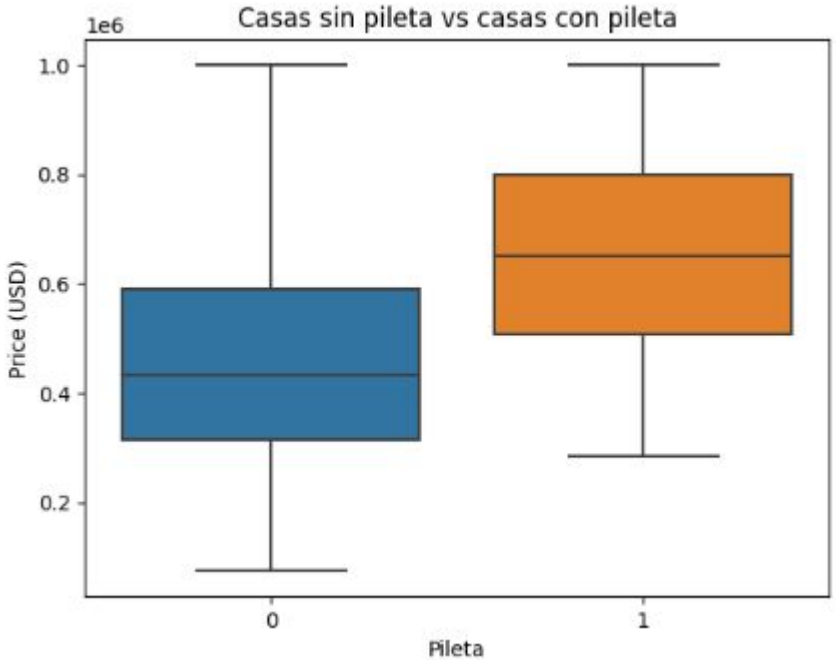
$433.000
Precio Mediana



Los datos fueron extraídos del siguiente notebook: link.

ANÁLISIS EXPLORATORIO

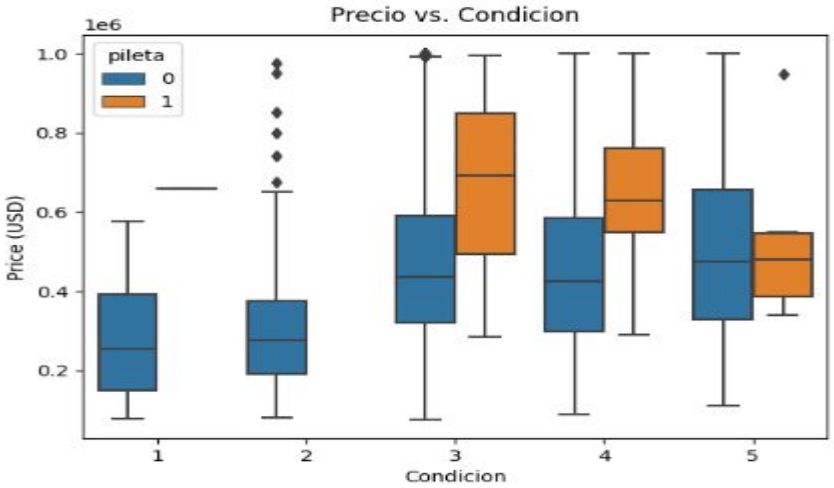
Hipótesis 1 - Piletas



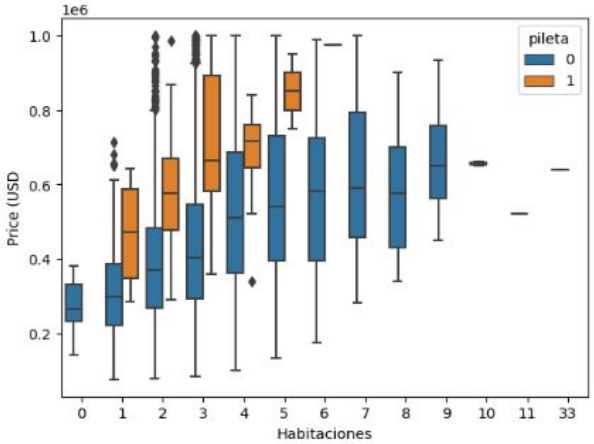
En primera instancia con un gráfico de boxplot vemos que las casas con pileta tienen una mediana mayor que las casas sin pileta

Por otro lado, vemos que se les asigna un valor de condición a las casas. Y podemos ver que las casas con pileta ya están en un rango mayor a 3. Pero solo la mediana es mayor para las casas en condición 3 y 4, por lo que podemos pensar que tal vez para ser una casa de rango 5 hay que tener en cuenta otros factores además de una pileta.

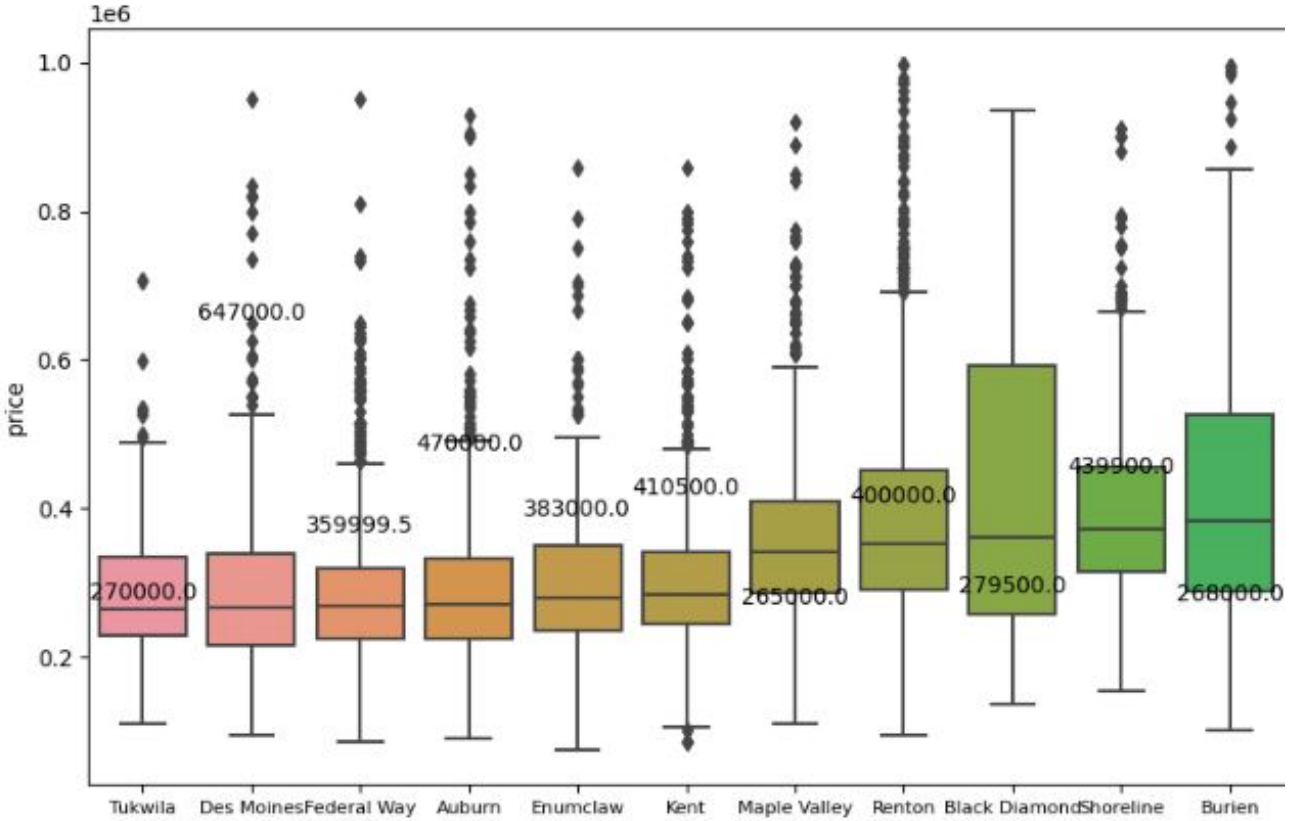
En el siguiente grafico vemos como las casas con hasta 6 ambientes del dataset tienen pileta y aumenta significativamente el precio en los casos que tienen pileta vs los que no tienen pileta.



Precio de Casas según cantidad de habitación con pileta y sin pileta



Hipótesis 2 - Zonas

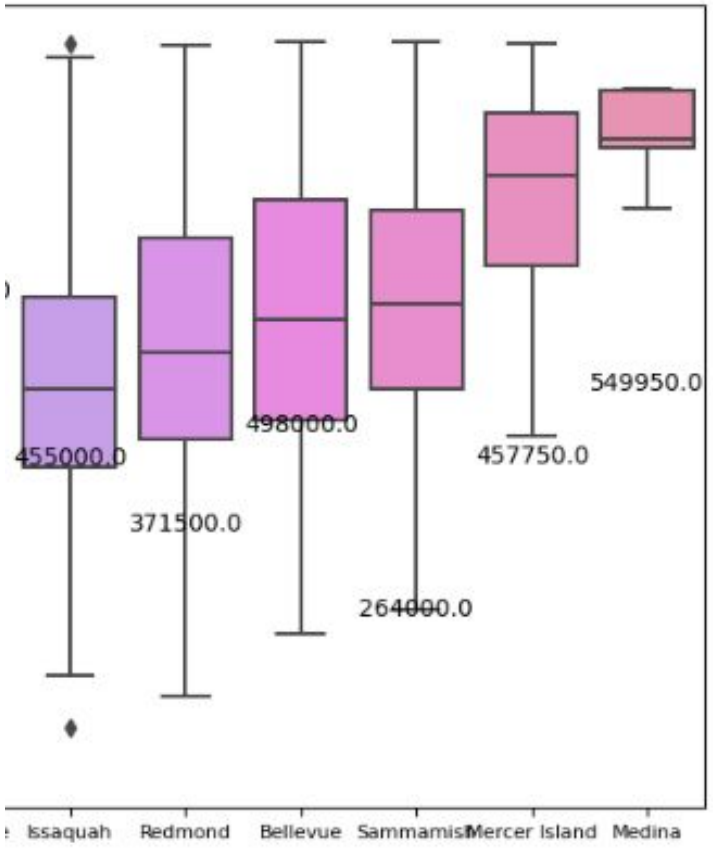


Agrupamos las zonas del dataset, y podemos analizar y mostrar en este gráfico las zonas con los valores más bajos del mercado.

Aun siendo las ciudades con medianas más bajas del estado se pueden observar outliers con casas de precios altos.

A continuación vemos las ciudades con las medianas más altas.

Y a diferencia del gráfico anterior aca no vemos outliers con precios bajos



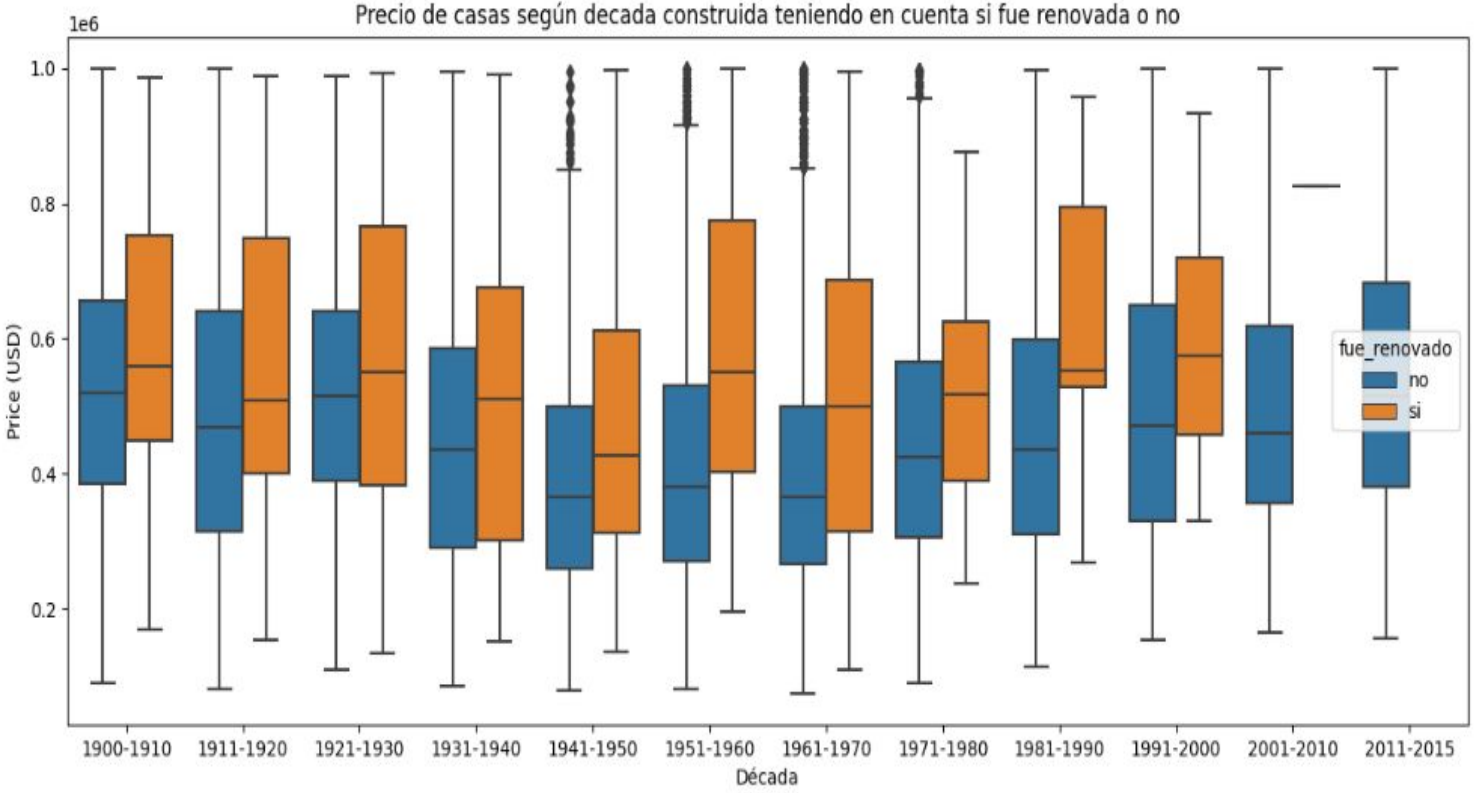
Hipótesis 3 - Año de Construcción

Ya vimos en el resumen del Meta Data el precio de las casas según la década.

Ahora agregamos la variable si fue renovada para ver si podemos obtener alguna nueva conclusión.

En principio la mediana de los precios aumenta para todos los casos, algo que es lógico. En ciertos casos los aumentos no son singificativos como es el caso de las casas construidas entre los años 1900 y 1930.

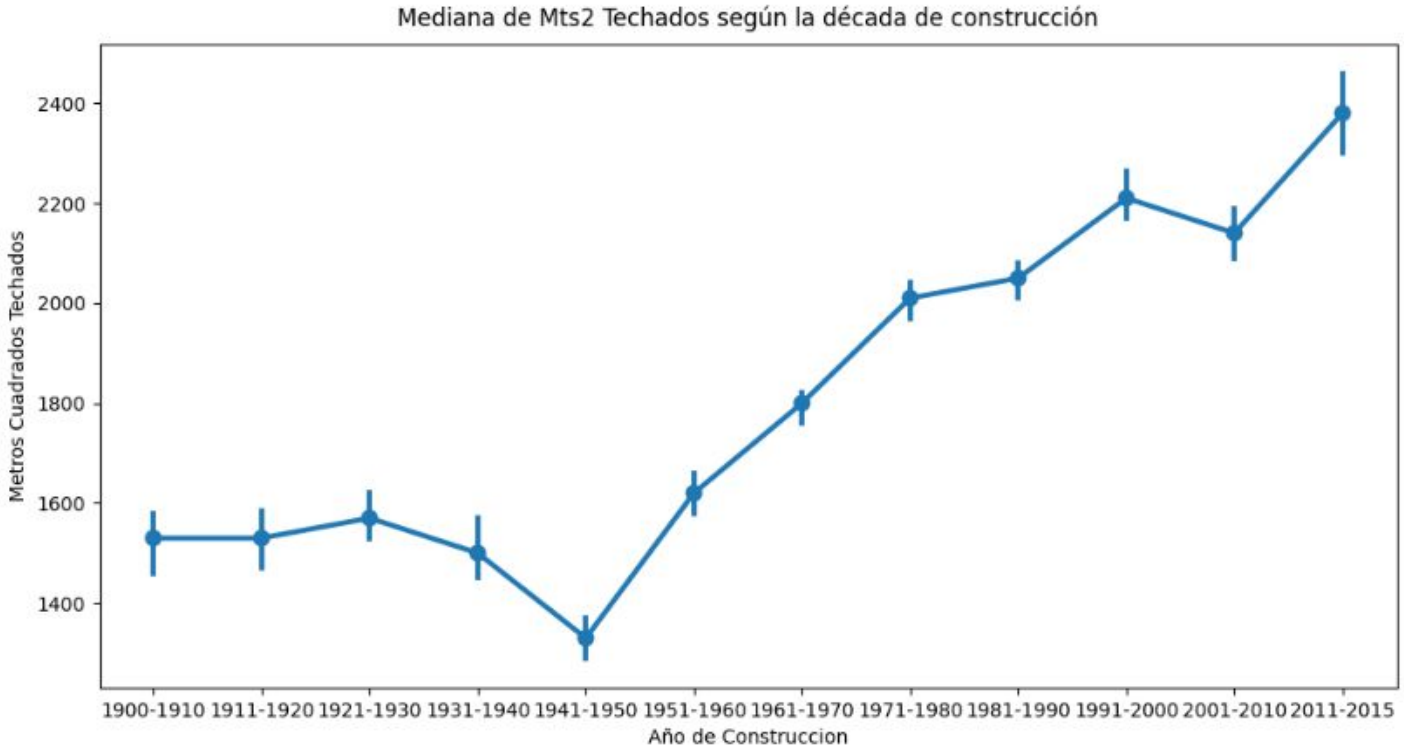
Pero por otro lado vemos que en las casas ubicadas en la década del 40, 50 y 60 los outliers de precios altos en el caso de que la casa no haya sido renovada desaparecen y pasan a ser parte del "bigote" del gráfico.



Para cerrar este análisis del año de construcción de las casas, comparamos con otra variable que no sea precio. Y vemos la mediana de los metros cuadrados de las casas según la década.

Se puede observar que a partir de la década del 70 inclusive las casas son cada vez más grandes.

En la última década casi que multiplican el tamaño de metros cuadrados techados con respecto a las casas de la década del 40. Por lo que puede ser una variable más que explica por qué las casas de la última década son las de un precio con mayor mediana.



INSIGHTS & **RECOMENDACIONES**

Conclusiones Hipótesis 1

Los precios de las casas tienden a aumentar en general, pero cuando se trata de casas con más de 5 o 6 habitaciones, el valor no aumenta proporcionalmente en comparación con las casas de tamaño más pequeño. Sin embargo, cuando se añade una pileta a la propiedad, se puede observar un aumento significativo del valor de la casa, incluso si tiene muchas habitaciones.

También hay otros factores que influyen cómo son los metros cuadrados techados y la condición en que se encuentra. En el caso de la condición vimos que en la condición de grado 5 (la mayor) no hace la diferencia tener la pileta.

En resumen, recomendaría la presencia de una pileta en una casa solo si son de tamaño pequeño o mediano puede aumentar significativamente su valor en el mercado inmobiliario. En una casa grande, seguramente son otras variables las que hacen atractiva a la casa.

Conclusiones Hipótesis 2

Al categorizar los códigos postales para cada zona que le corresponde podemos ver como hay 2 zonas con los precios con la mediana más alta y destacada del resto que son Island y Medina con lo que podríamos decir que son las zonas más caras y exclusivas del estado.

Luego podemos ver que Tukwila, Des Moines, Federal, Auburn y Enumclaw son las zonas con los precio más bajos del estado de Washington.

Sin embargo a excepción de Tuwkila, todas las ciudades tienen outliers que llegan a estar al precio de las ciudades más caras. Por lo que podemos decir que en mayor o menor medida, existen casas con precios muy altos en casi todas las ciudades.

Conclusiones Hipótesis 3

- Las casas más antiguas, construidas entre 1900 y 1930, y las casas de la última década tienen precios más altos en comparación con otras décadas.

- La mediana de los precios aumenta en general, pero este aumento no es significativo para las casas construidas entre 1900 y 1930. Además, se observa que en las décadas del 40, 50 y 60, los precios más altos se encuentran principalmente en las casas que no fueron renovadas.

- Las casas más recientes tienen un tamaño de metros cuadrados techados mucho mayor en comparación con las casas más antiguas. Este aumento en el tamaño de las casas puede ser un factor adicional que explica por qué las casas de la última década tienen precios más altos. Siguiendo el mismo análisis que la hipótesis número 1.

En general, estos hallazgos sugieren que el año de construcción y el tamaño de la casa son variables importantes a considerar al analizar los precios de las casas. Las casas más antiguas y las más recientes tienden a tener precios más altos, y el tamaño de la casa también juega un papel significativo en la determinación del precio.

Feature Selection

El dataset estaba no contaba con valores nulos pero algunas variables necesitaron ser cambiada de tipo para realizar un mejor análisis.

```
# Ajustamos las variables del dataset para poder realizar el feature selection correctamente
df_casas['baños'] = df_casas['baños'].str.replace(',', '.').astype(float)
df_casas['pisos'] = df_casas['pisos'].str.replace(',', '.').astype(float)
df_casas['fue_renovado'] = df_casas['fue_renovado'].map({'si': 1, 'no': 0})
df_casas['fue_renovado'] = df_casas['fue_renovado'].astype(int)
```

Luego realizamos el feature selection para obtener las variables con mejor correlación con la variable a predecir.
A partir de esto podemos analizar los modelos y elegir el de mejor performance.

Este trabajo lo realizamos dos veces ya que intenté agregar 3 variables nuevas para ver si encontraba mejor correlación pero me quede con el primer feature ya que tuvo mejor resultado.

Evaluación de Modelos

Probe 3 algoritmos con los 2 feature selection:

- Regresión Lineal Múltiple
- Regresión Lineal Simple
- XGBOOST

El modelo de mejor performance fue XGBOOST con el primer feature selection. Siendo los siguientes los resultados obtenidos:

```
Mean Absolute Error (MAE): 56,394.11
Mean Squared Error (MSE): 6,180,463,655.03
Root Mean Squared Error (RMSE): 78,615.92
```

A continuación dejo un gráfico donde comparo los datos predecidos vs los datos reales con XGBOOST