



# BUSINESS CASE

Caso práctico de estimación de precios de inmuebles



# CONTENIDO

- 01 INTRODUCCION Y MARCO TEORICO
- 02 PLANTEAMIENTO DEL PROBLEMA
- 03 METODOLOGÍA
- 12 RESULTADOS
- 13 CONCLUSIONES
- 14 ANEXOS

# INTRODUCCION Y MARCO TEORICO

## Proyecto de predicción de precios de casas

En el mercado inmobiliario, el precio de los inmuebles está determinado por una variedad de factores que van más allá de la mera superficie de la propiedad. Factores como la ubicación, tamaño de la vivienda y otras características específicas pueden influir sobre la determinación del inmueble. Por ello, la predicción de los precios de los inmuebles se ha convertido en una herramienta valiosas para los compradores, como los vendedores.

Este estudio se basa en un conjunto de datos que contiene información detallada sobre **21,631** viviendas en el estado de Washington. El objetivo principal es construir un modelo de Machine Learning capaz de predecir el precio de una vivienda en función de sus características. La predicción precisa de los precios puede tener un impacto positivo, no solo en la toma de decisiones en el mercado inmobiliario, sino también en la valorización y optimización de los recursos disponibles.



# PLANTEAMIENTO DEL PROBLEMA

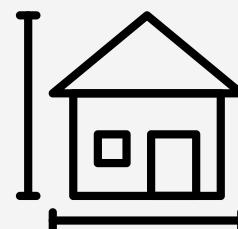
## Predicción de precio de inmuebles

Tal como se mencionó con anterioridad, la información proporcionada corresponde al estado de Washington, donde se cuentan con un total de **21,631** observaciones. La idea es realizar análisis, procesamiento de variables y un modelo predictivo que se encargue en estimar el precio de la vivienda basándose en las 21 variables iniciales presentadas, dentro de las cuales se consideran:



### Información del inmueble (3 variables):

- ID del inmueble
- Fecha de construcción
- Código postal



### Medidas del inmueble (6 variables):

- Tamaño del inmueble
- Tamaño sótano

### Características del inmueble (9 variables):

- Número de habitaciones
- Número de pisos



### Características demográficas (2 variables):

- Longitud
- Latitud

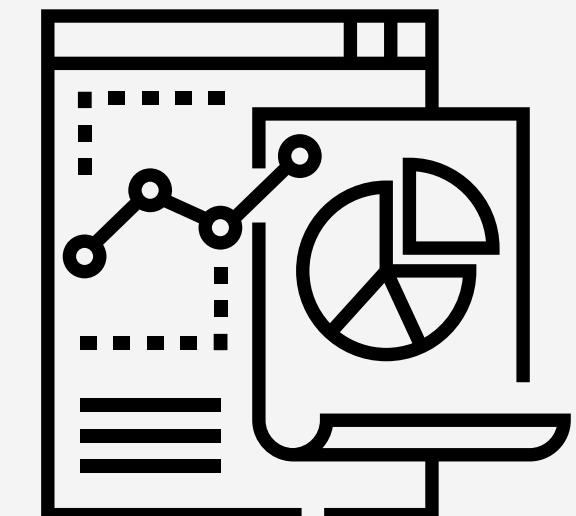


# METODOLOGIA

## Análisis de los datos

Realizar un análisis exploratorio de los datos (EDA) es crucial en una elaboración del modelo. No solo ayuda a visualizar la variable dependiente, si no que proporciona información sobre la relación de las variables y su distribución. Dentro de este análisis, se realizó el siguiente proceso:

- Valoración y resumen estadístico de variables
- Exploración gráfica de variables
- Análisis de correlación de variables



# METODOLOGIA

## Valoración y resumen estadístico de variables

Dentro de esta valoración, se encontró:

- De las 21 variables iniciales:
  - No hay valores nulos ni repetidos
  - 7 variables son categóricas. (Número de habitaciones, baños, pisos, vista, ventanas, condición y grado)
  - 14 variables son numéricas.
- Los resultados estadísticos mostraron:
  - Un rango de precios de \$75,000 y \$7,700,000. Con un promedio de \$540,088
  - 3 habitaciones y 2 baños en promedio, con inmuebles de 1 a 2 pisos promedio.
  - Tamaño promedio de 2,079 pies.
  - Los resultados de latitud y longitud muestran que los inmuebles están en una misma ubicación (estado).
  - Solo hay una variable booleana (0,1) siendo waterfront y solo 163 tienen este valor positivo.

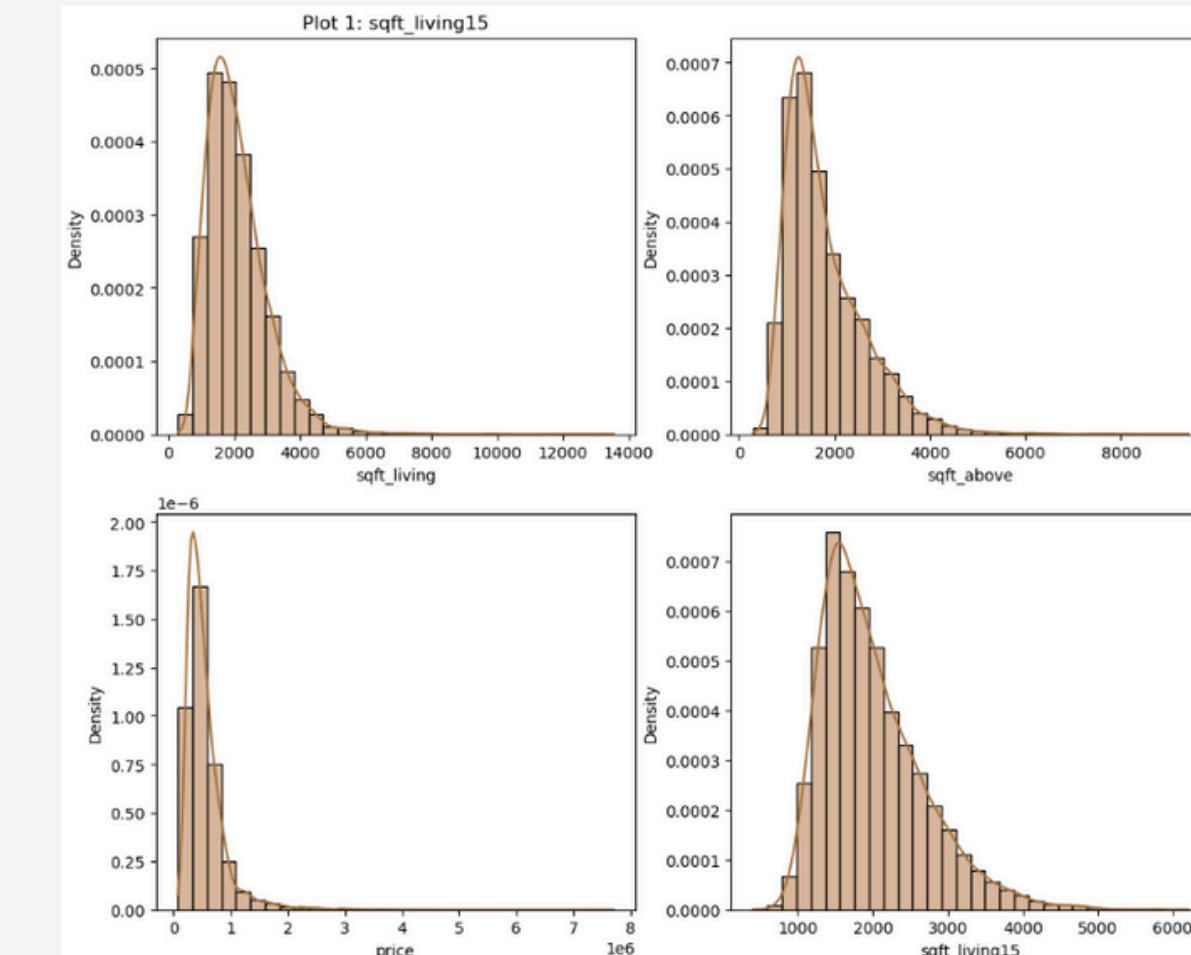
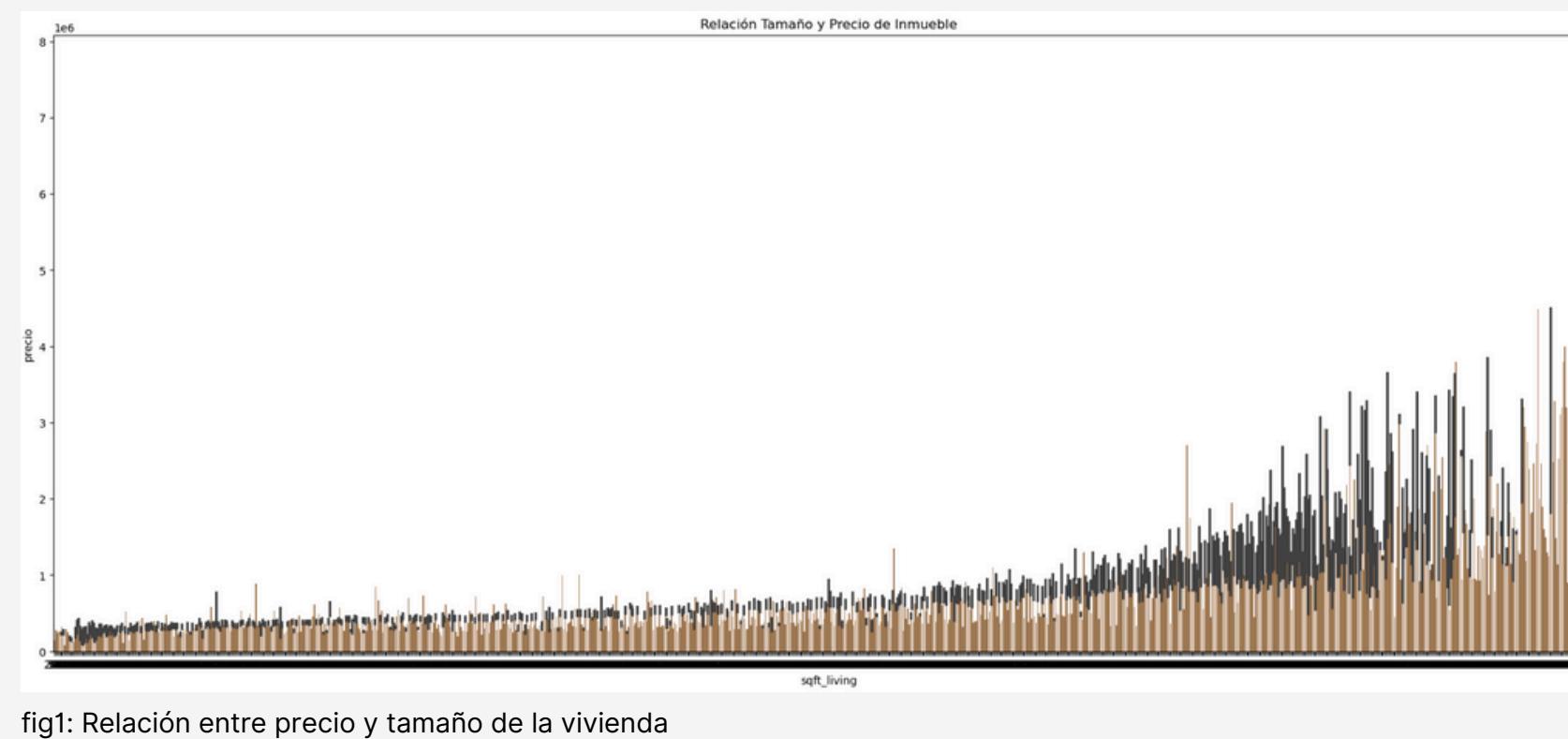


# METODOLOGIA

## Exploración gráfica de variables y correlación

Realizar un análisis gráfico sobre las variables independientes y la variable dependiente (precio), ayudará a visualizar los datos y verificar si siguen una distribución. Dichos análisis partieron de revisar histogramas, densidades y pairplots (gráficas 1 a 1 por cada variable). Los resultados más relevantes fueron:

- Existe una gran relación entre el precio del inmueble y el tamaño de la vivienda (correlación = .7)
- Las variables relacionadas al tamaño medido en pies muestran una acumulación en la media.

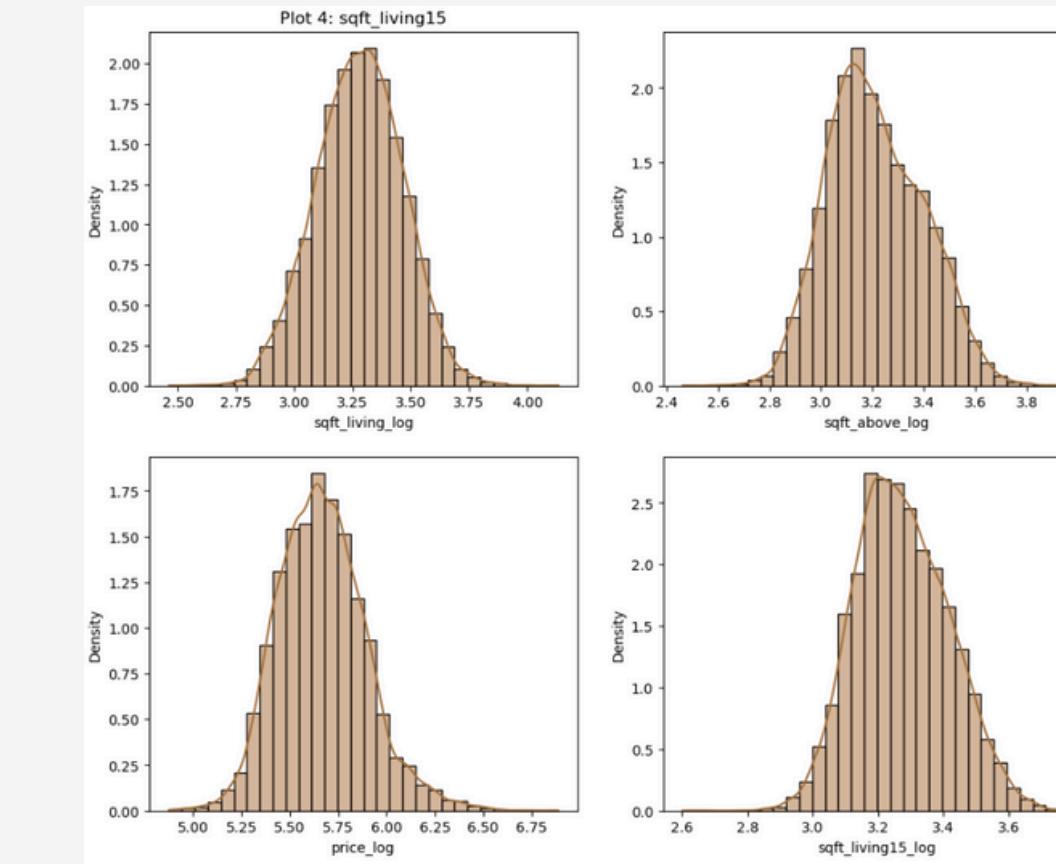
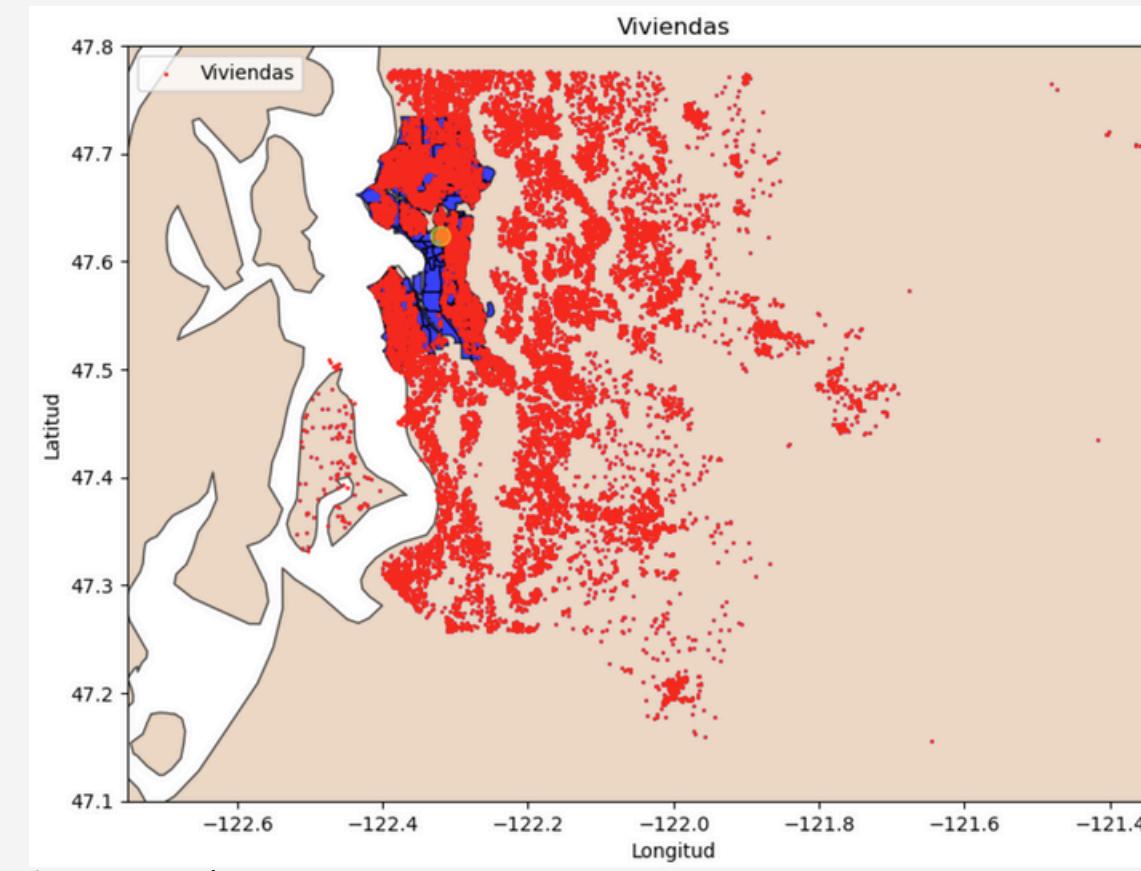


# METODOLOGIA

## Procesamiento de variables

Utilizando las observaciones anteriores, se pudieron transformar y procesar nuevas variables. Las cuales fueron:

- Transformación logarítmica de variables: Dado que estas variables tienen una distribución donde reside gran acumulación, al transformar los valores numéricos, podría facilitar la normalización de los datos.
  - Precio (price)
  - Tamaño de vivienda (sqft\_living)
  - Tamaño de sótano (sqft\_above)
  - Tamaño de vivienda transformada (sqft\_living15)
- Adición de distancia: Se generó una variable que mide la distancia entre cada vivienda y el centro de la zona con mayor plusvalía en Seattle (Capitol Hill).



# METODOLOGIA

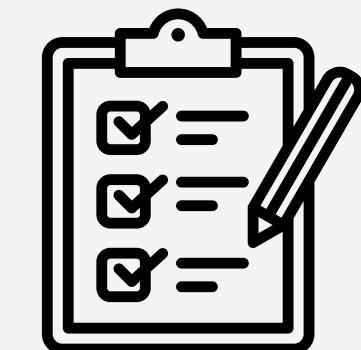
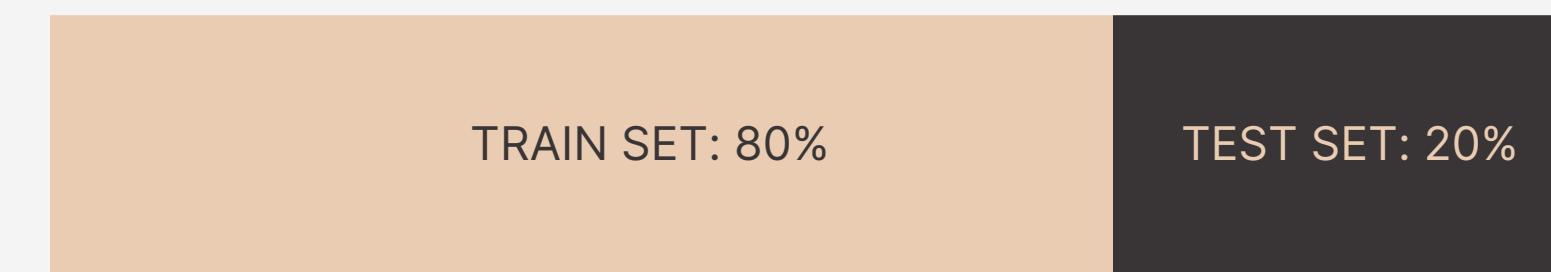
## Incorporación de modelos

Dado que ya se cuenta con el dataset final, siendo de 20 variables finales, y teniendo en consideración la variable dependiente **price** se procede a proponer los modelos predictivos.

- Modelo de regresión lineal: Este modelo asumirá la relación lineal entre la variable dependiente y una sola variable independiente.
- Modelo de regresión múltiple: A diferencia de la regresión lineal, este modelo extiende la regresión lineal incorporando múltiples variables independientes.
- Modelo de Regresión Polinomial: Por último, el modelo de regresión polinomial además de incorporar el resto de variables, ajustará los comportamientos no lineales.

Al final, cada uno de los modelos predictivos serán evaluados utilizando métricas de rendimiento, siendo:

- R<sup>2</sup>: Esta métrica nos da una idea general de qué tan bien se ajusta el modelo.
- MSE (Error Cuadrático Medio): Calcula el promedio de los cuadrados de los errores entre las predicciones del modelo y los valores reales observados
- MAE (Error Absoluto Medio): Mide el error promedio en términos absolutos, a diferencia del MSE, este no penaliza los errores grandes de manera tan fuerte.



# METODOLOGIA

## Modelo 1 - Regresión lineal.

Con el fin de realizar distintas versiones de un modelo de regresión lineal. Se propusieron 4 distintos modelos, considerando la variable independiente como las 4 con las que tuviera más relación el precio.

- Modelo 1: Precio, Grado (correlación .7) [fig5]
- Modelo 2: Precio, Medida de vivienda (correlación de .67) [fig6]
- Modelo 3: Precio, Medida de sótano (correlación de .59) [fig7]
- Modelo 4: Precio, Medida de vivienda transformada (correlación .61) [fig8]

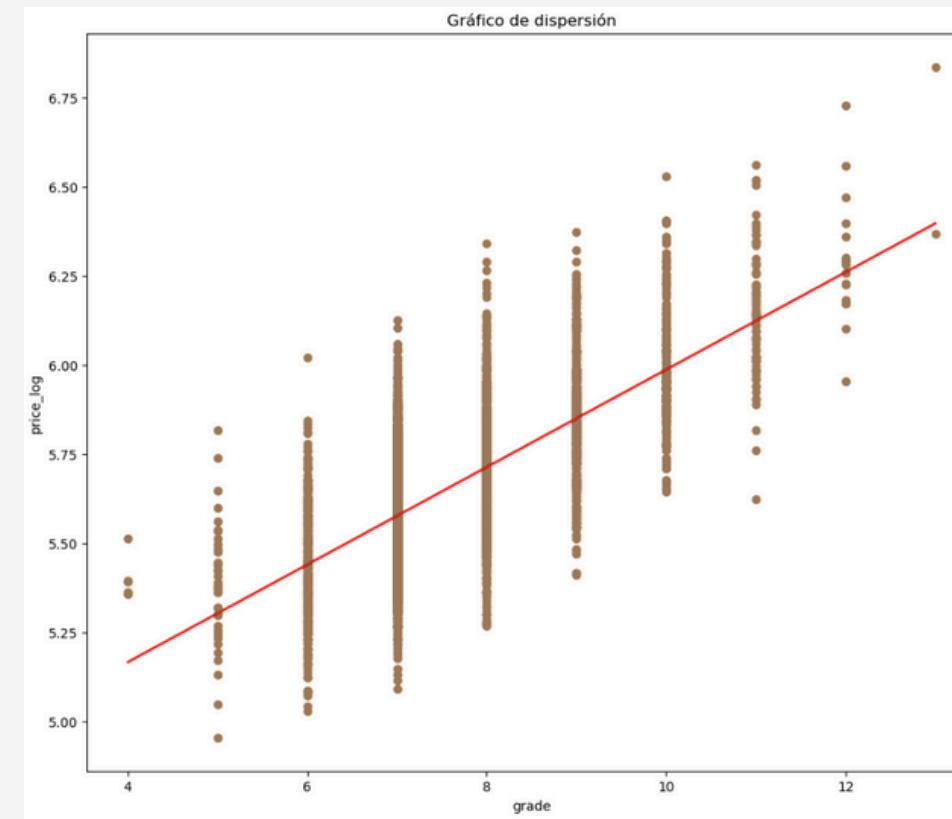


fig5: Gráfico de dispersión (price\_log, grade)

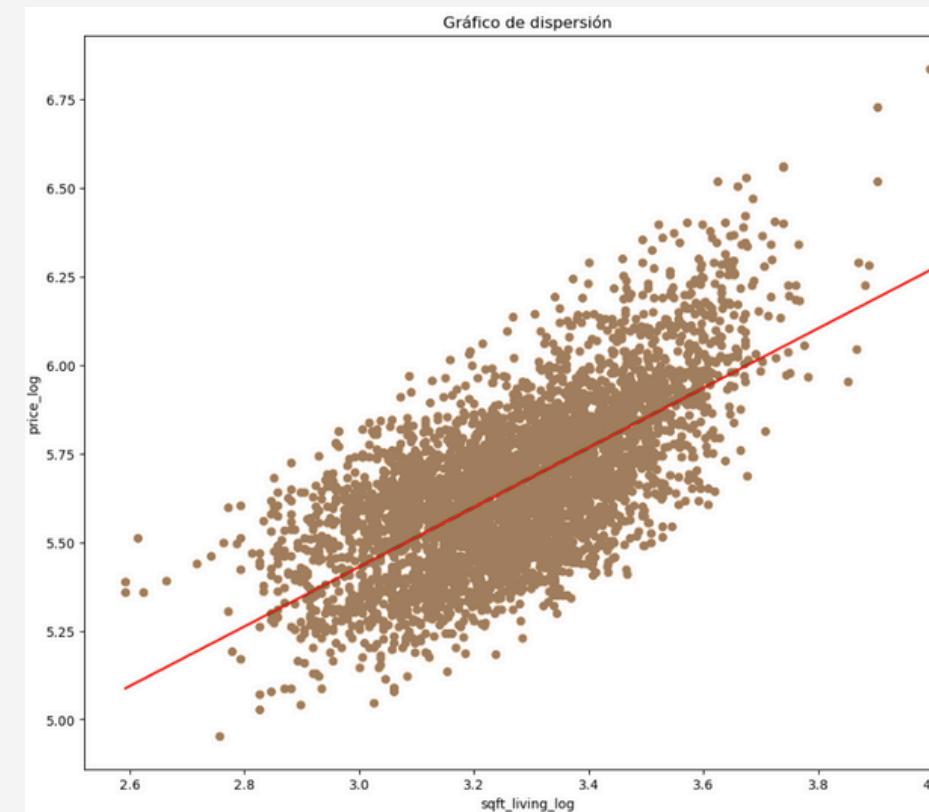


fig6: Gráfico de dispersión (price\_log, medida de vivienda)

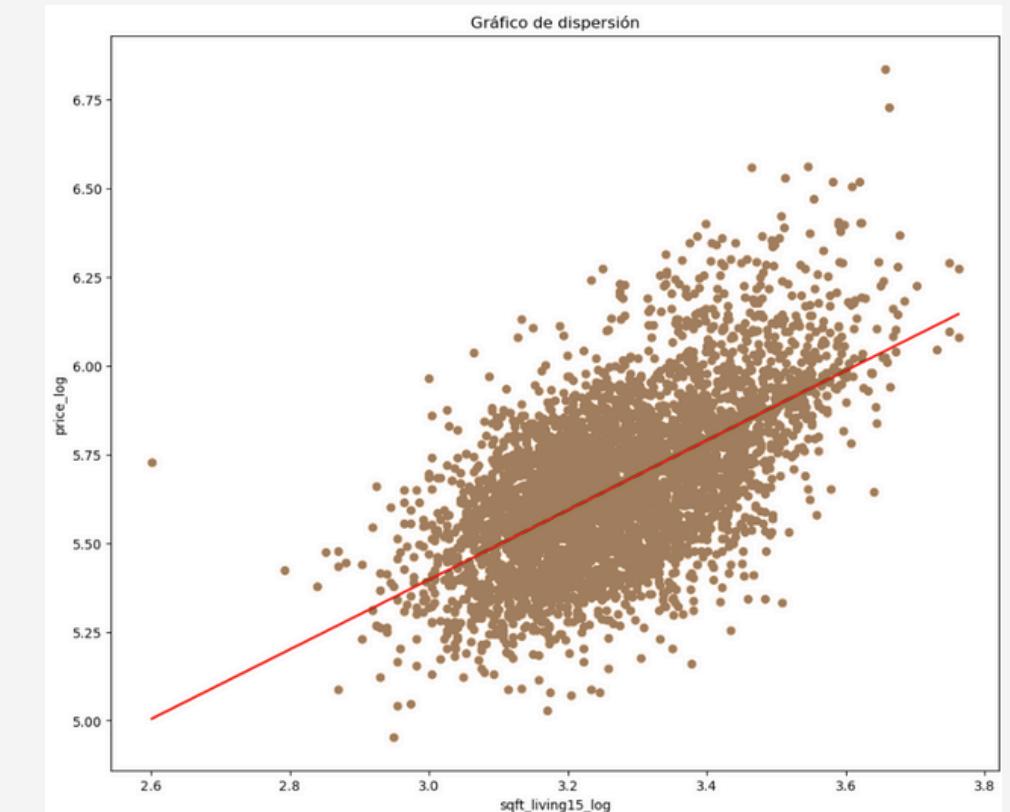


fig8: Gráfico de dispersión (precio, medida de vivienda transformada)

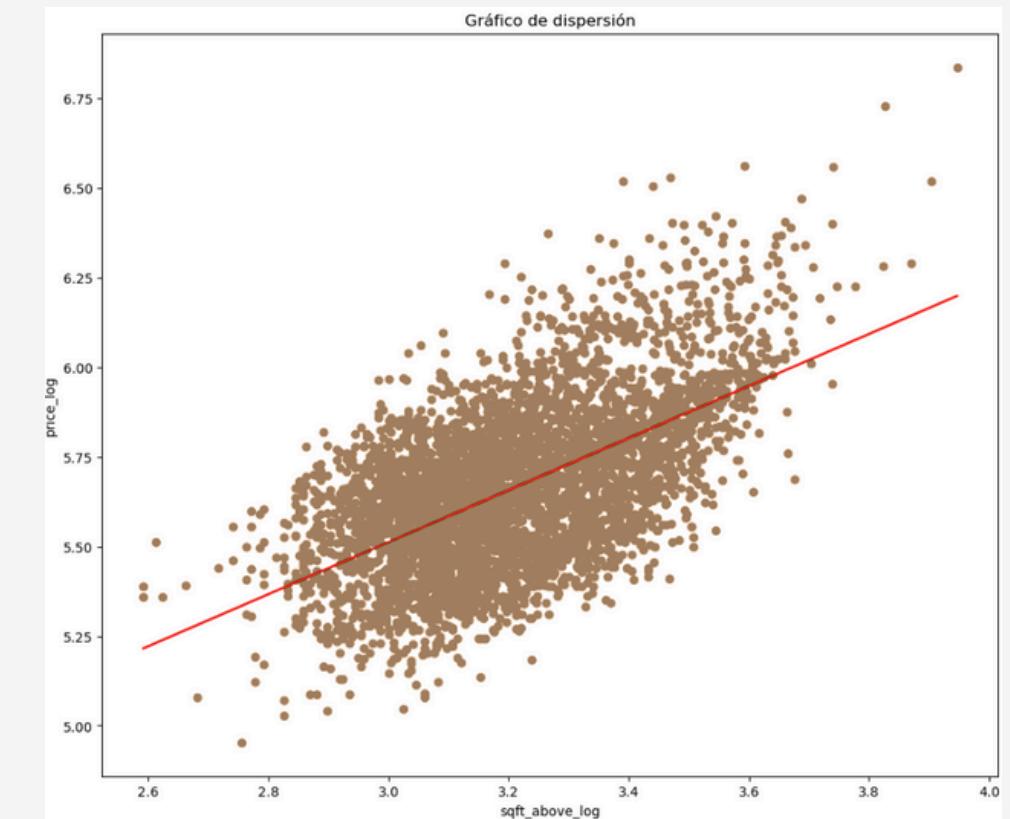


fig7: Gráfico de dispersión (precio, medida de sótano)

# METODOLOGIA

## Modelo 1 - Regresión lineal.

Modelo	R <sup>2</sup>	MSE	MAE
LRM (price, grade)	48.5107%	2.62882%	12.94832%
LRM(price, sqft_living)	44.73603%	2.8215%	13.6656%
LRM(price, sqft_above)	33.6773%	3.3861%	14.8356%
LRM(price, sqft_living15)	35.8988%	3.2727%	14.2737%

# METODOLOGIA

## Modelo 2 - Regresión múltiple.

Similar al caso de regresión lineal, se realizaron 2 modelos de regresión múltiple

- Regresión múltiple considerando todas las variables[fig9]
  - $R^2 = 80.809\%$
  - $MSE = .97979\%$
  - $MAE = 7.6827\%$
- Regresión múltiple considerando 10 variables más relacionadas con **price** [fig10].
  - $R^2 = 72.0747\%$
  - $MSE = 1.425\%$
  - $MAE = 9.303\%$

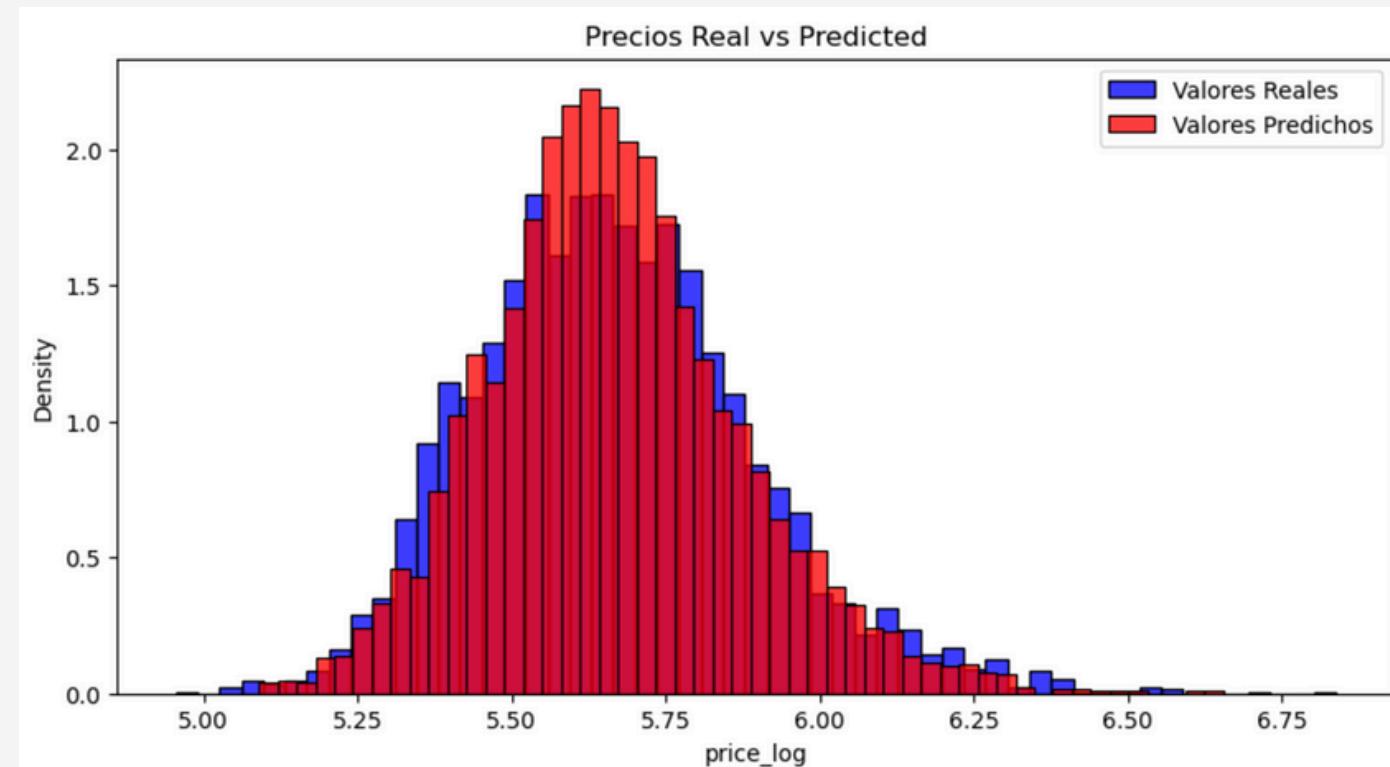


fig9. Valores predichos y reales, considerando todas las variables

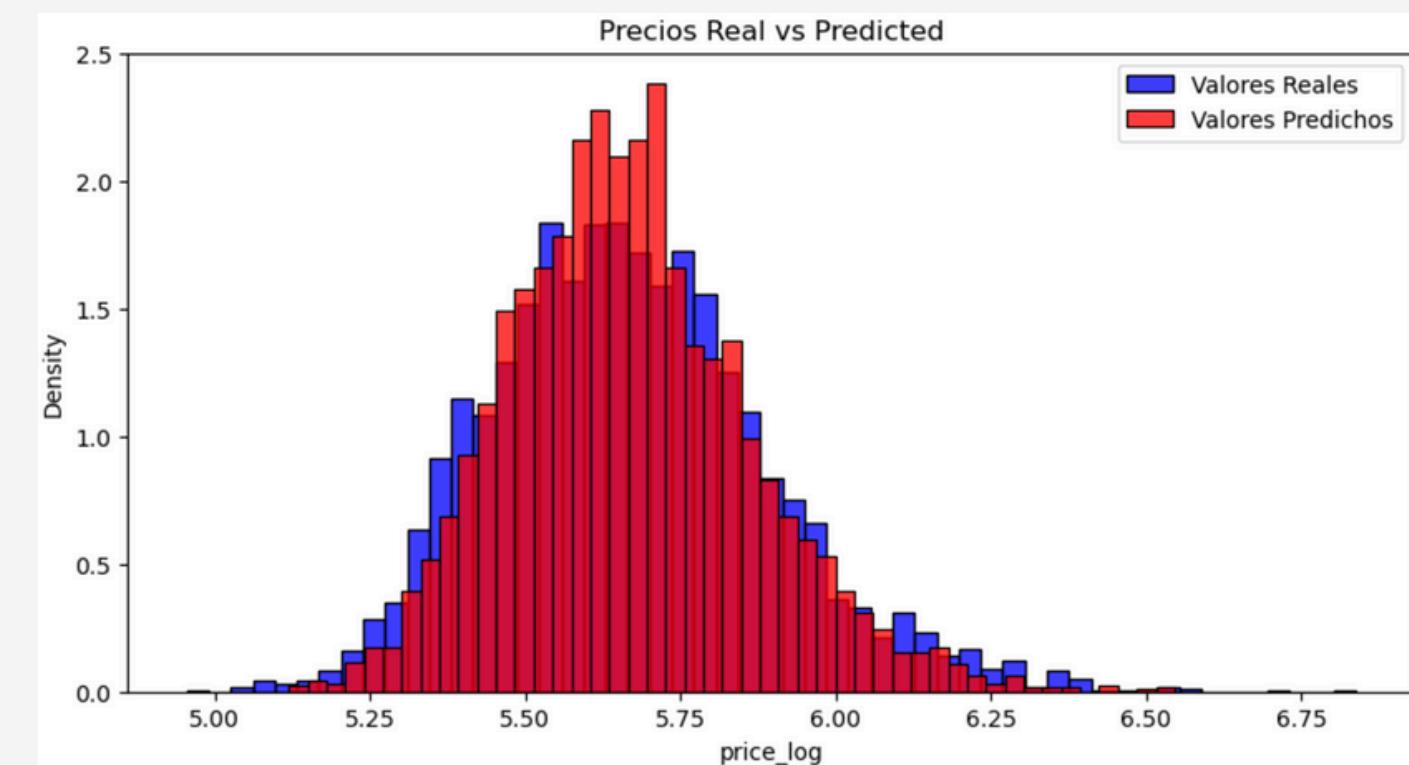


fig10. Valores predichos y reales, considerando 10 variables más relacionadas

# METODOLOGIA

## Modelo 3 - Regresión Polinomial.

Similar al caso de regresión múltiple, se realizaron 2 modelos de regresión polinomial.

- Regresión polinomial considerando todas las variables[fig10]
  - $R^2 = 86.525\%$
  - $MSE = .6879\%$
  - $MAE = 6.075\%$
- Regresión polinomial considerando 10 variables más relacionadas con **price** [fig11].
  - $R^2 = 75.228\%$
  - $MSE = 1.2647\%$
  - $MAE = 8.6573\%$

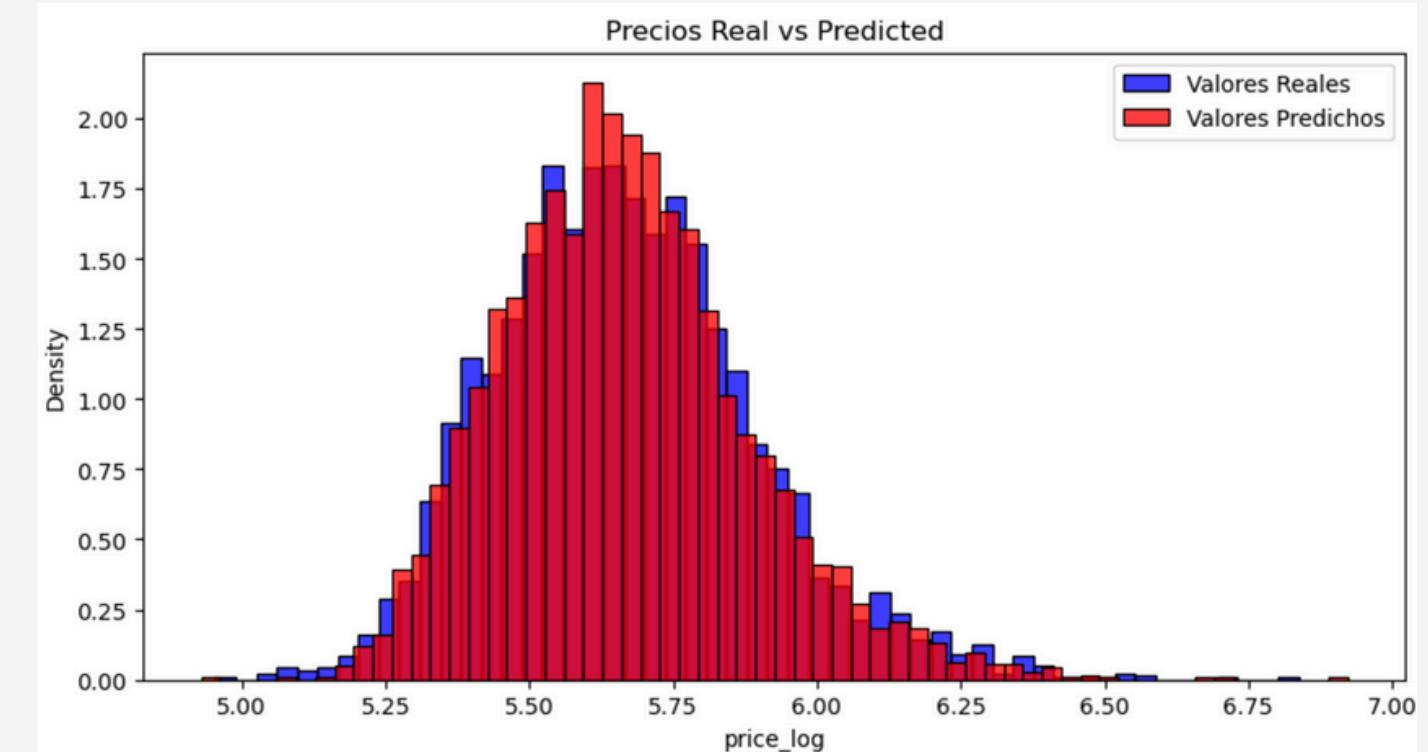


fig11. Valores predichos y reales, considerando todas las variables

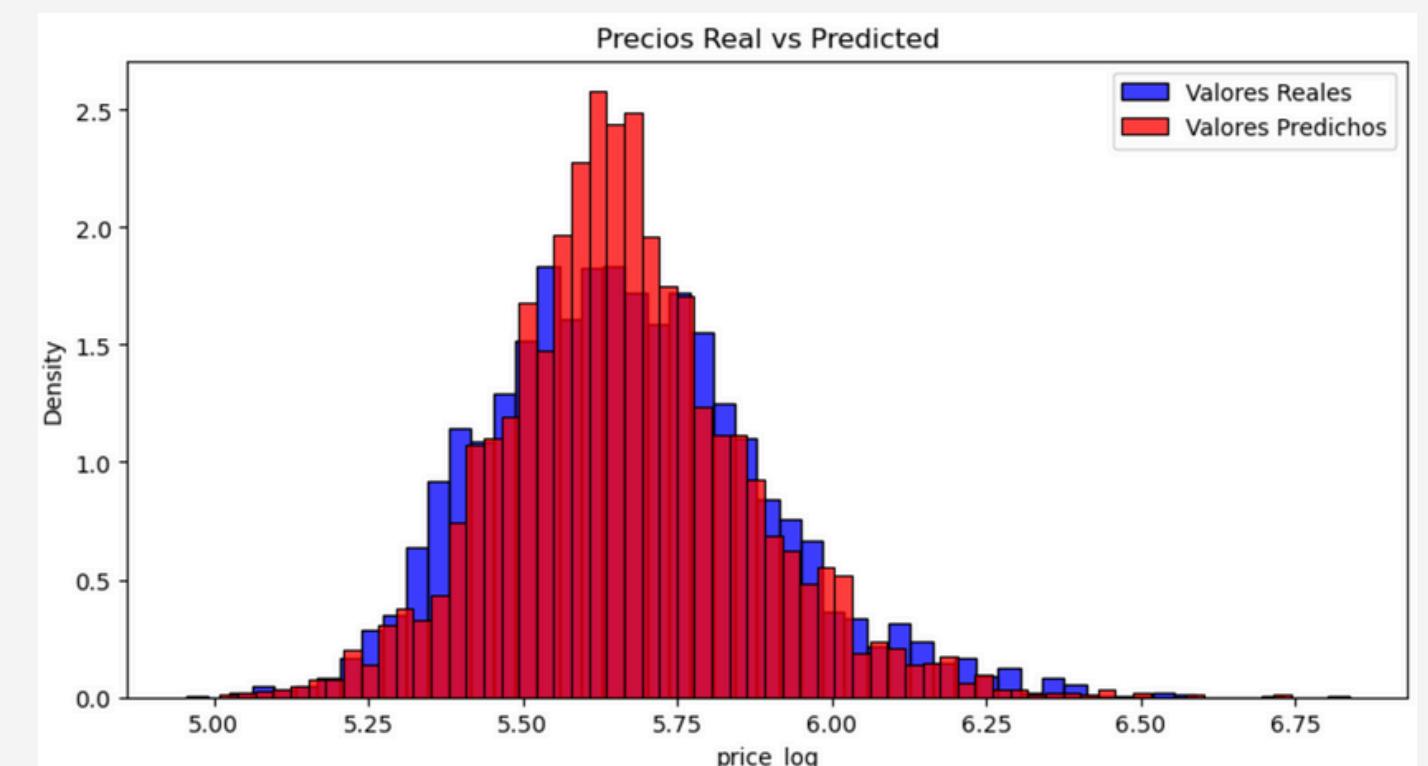


fig12. Valores predichos y reales, considerando 10 variables más relacionadas

# RESULTADOS

Modelo	Set	R <sup>2</sup>	MSE	MAE
LRM (price, grade)	Test	48.5107%	2.62882%	12.94832%
LRM (price, grade)	Train	49.7524%	2.6446%	12.9811%
LRM (price, sqft_living)	Test	44.73603%	2.8215%	13.6656%
LRM (price, sqft_living)	Train	45.7446%	2.8556%	13.669%
LRM (price, sqft_above)	Test	33.6773%	3.3861%	14.8356%
LRM (price, sqft_above)	Train	34.5413%	3.4453%	14.937%
LRM (price, sqft_living15)	Test	35.8988%	3.2727%	14.2737%
LRM (price, sqft_living15)	Train	37.1006%	3.106%	14.3761%

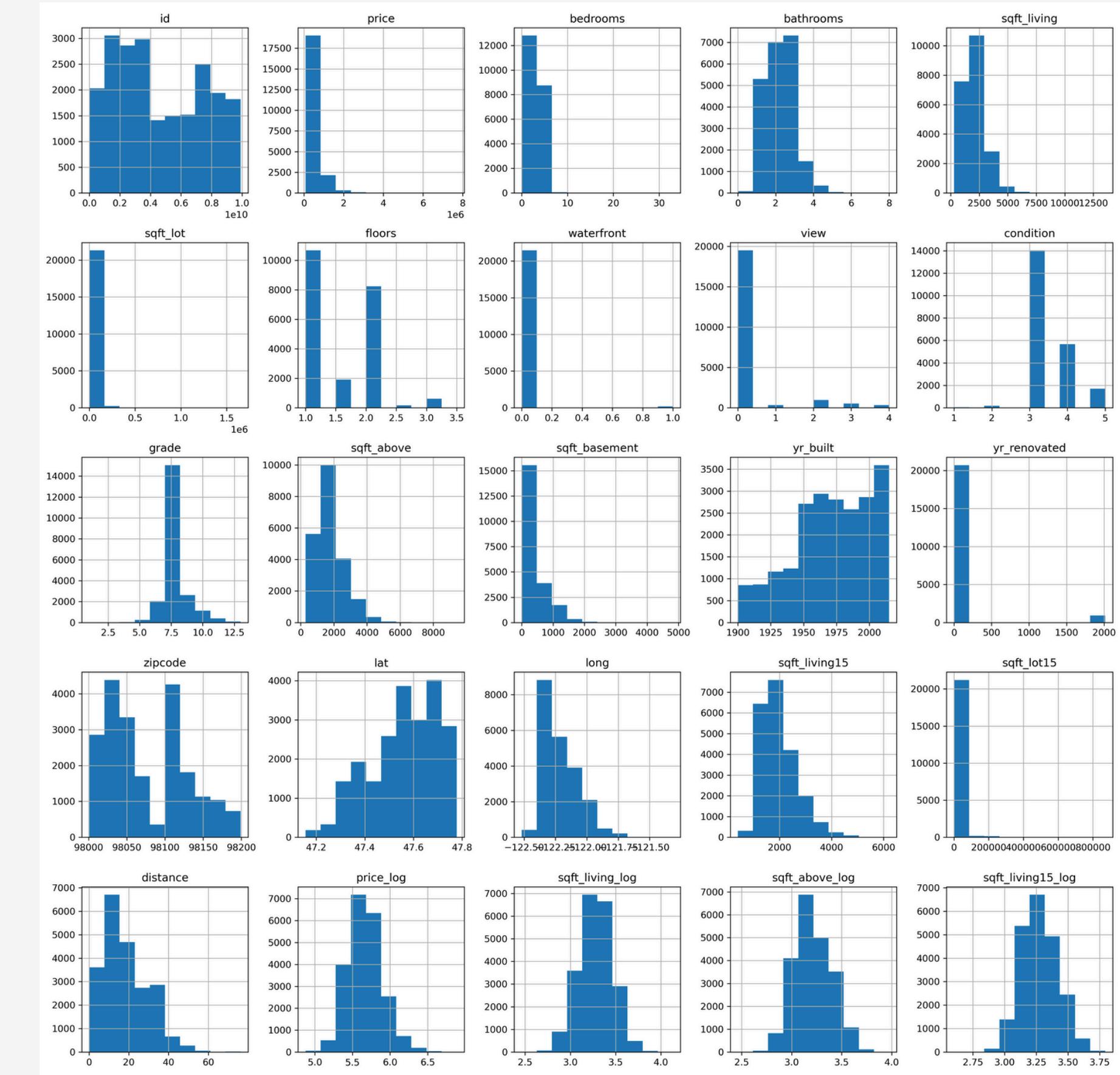
Modelo	Set	R	MSE	MAE
MR (All vars)	Test	80.809%	.97979%	7.6827%
MR (All vars)	Train	81.70856%	.9627%	7.5474%
MR (All vars)	CV	73.02%	1.4%	9.18%
MR (10 vars)	Test	72.0747%	1.425%	9.303%
MR (10 vars)	Train	73.4945%	1.395%	9.135%
PR (All vars)	Test	86.5254%	.6879%	6.607%
PR (All vars)	Train	88.2041%	.6208%	5.887%
PR (All vars)	CV	76.18%	1.24%	8.54%
PR (10 vars)	Test	75.2283%	1.2647%	8.6573%
PR (10 vars)	Train	76.9008%	1.2157%	8.447%

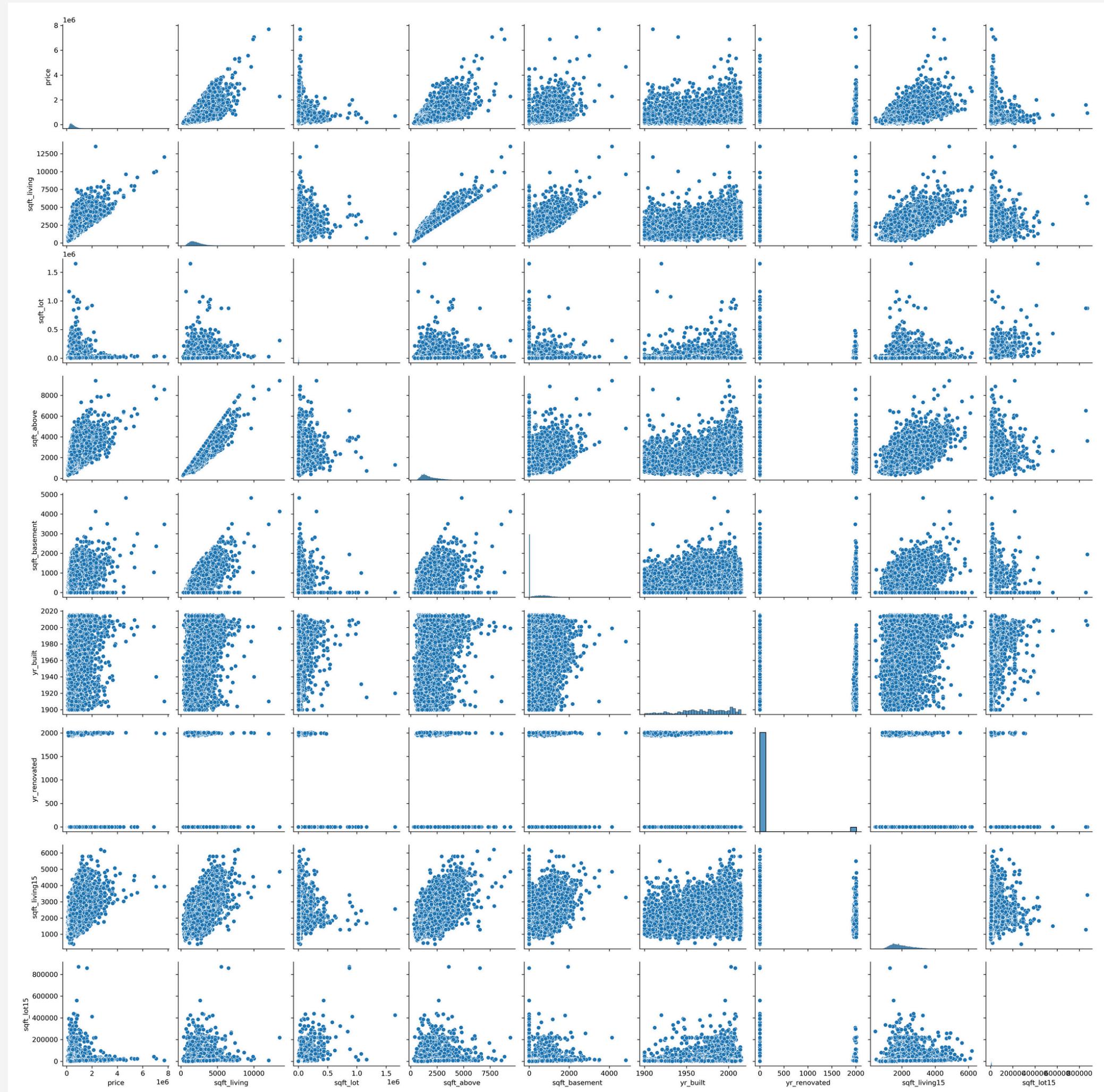
# CONCLUSIONES

Tras haber analizado, trabajado y evaluado los modelos utilizados para la predicción de precio de inmuebles:

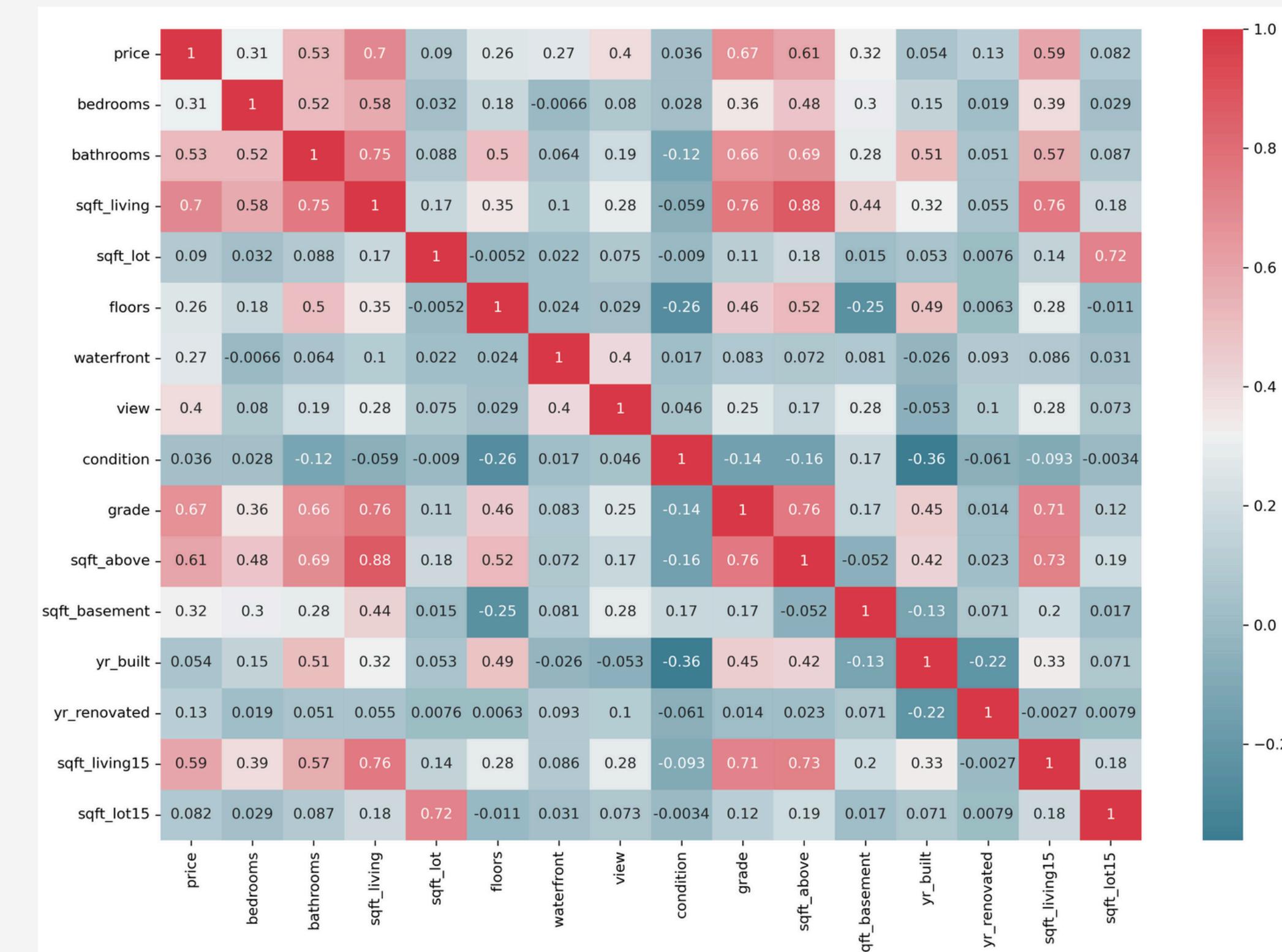
- En general, ningún modelo presentó una estimación incorrecta
- Existe una gran relación entre el precio del inmueble y las medidas del mismo
- Se suele pensar que una regresión polinomial ajusta de manera ‘forzada’ los datos, sin embargo, en caso de no optar por este modelo, está la segunda opción de utilizar una regresión múltiple.
- Utilizar un grado mayor a 2 podría repercutir negativamente en la estimación del modelo (en el caso de regresión polinomial).
- A pesar que la distancia no presenta una relación muy directa, si se trabajara mejor sobre dicha variable, tal vez tomar en consideración más barrios de plusvalía, se podría aumentar la relación.
- Utilizar Cross-Validation suele ser una buena alternativa, pues pruebaé una robustez en la predicción de los datos, además de un acercamiento más real de los datos.

# ANEXOS

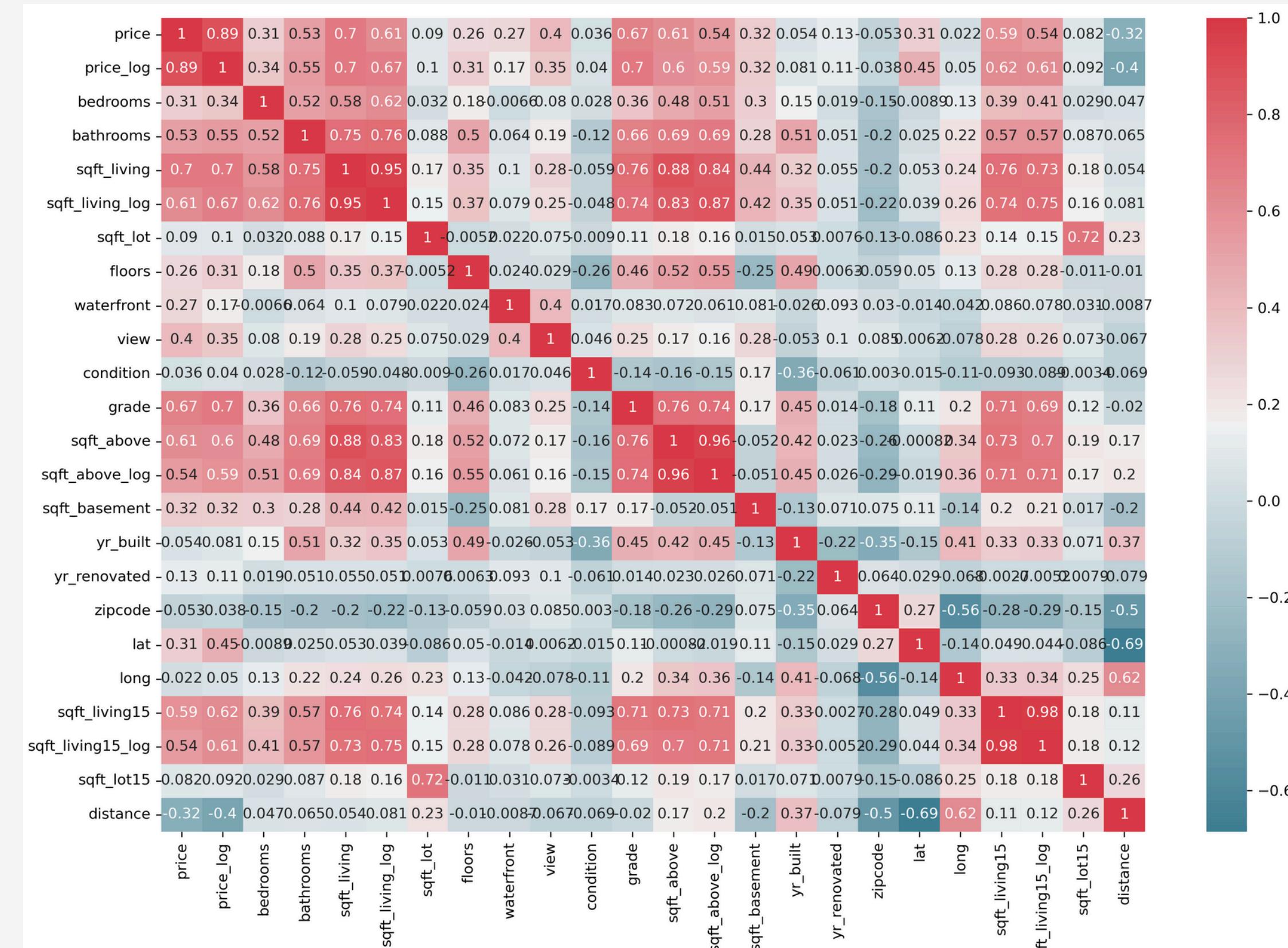




## Mapa de correlación inicial



## Mapa de correlación final



## Predicciones para Regresión múltiple y Polinomial con Cross-Validation

