Predicción valores de renta de alojamientos turisticos en Barcelona, España

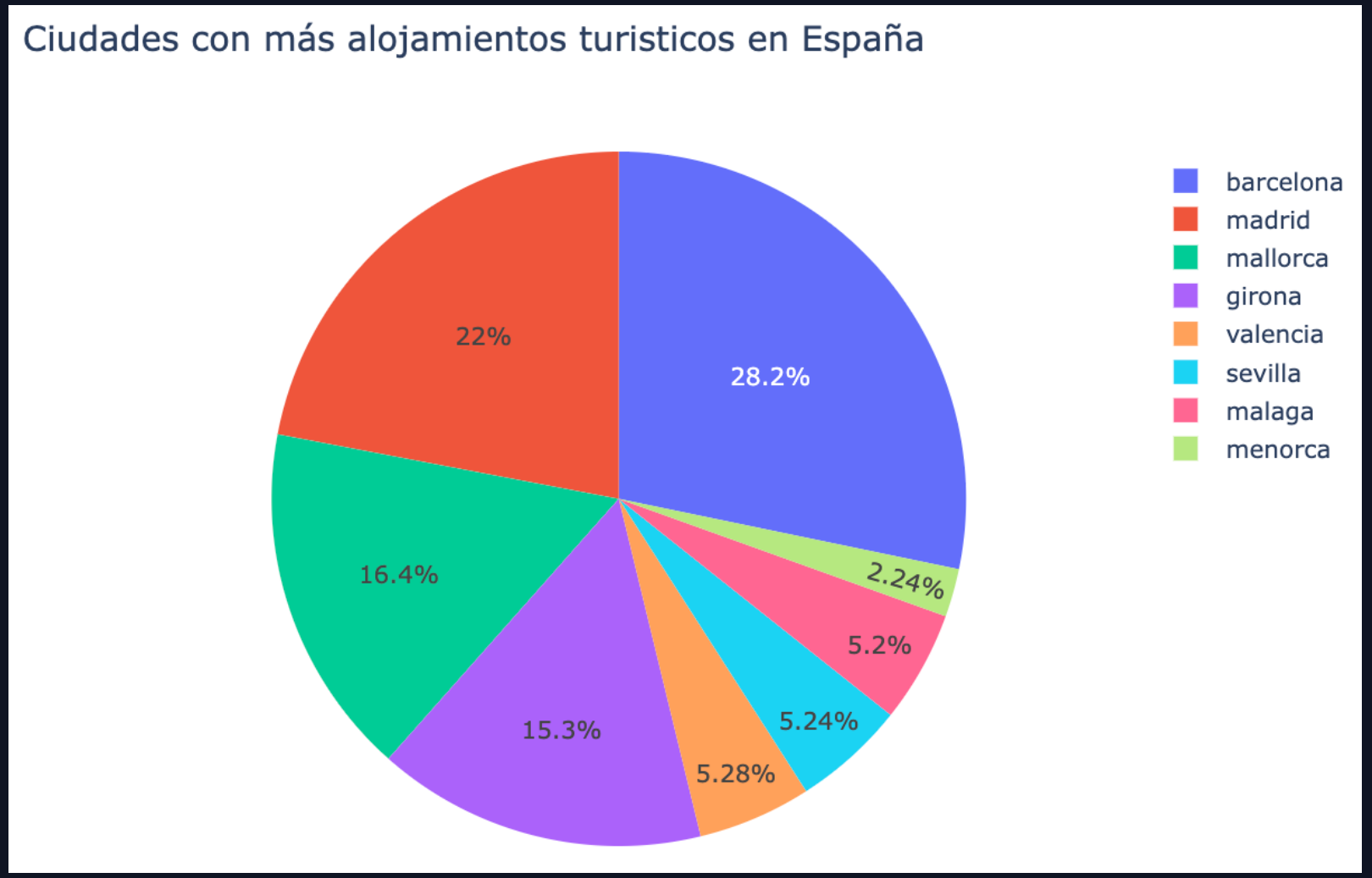
Cristian Vega Castro

Alojamientos en Barcelona

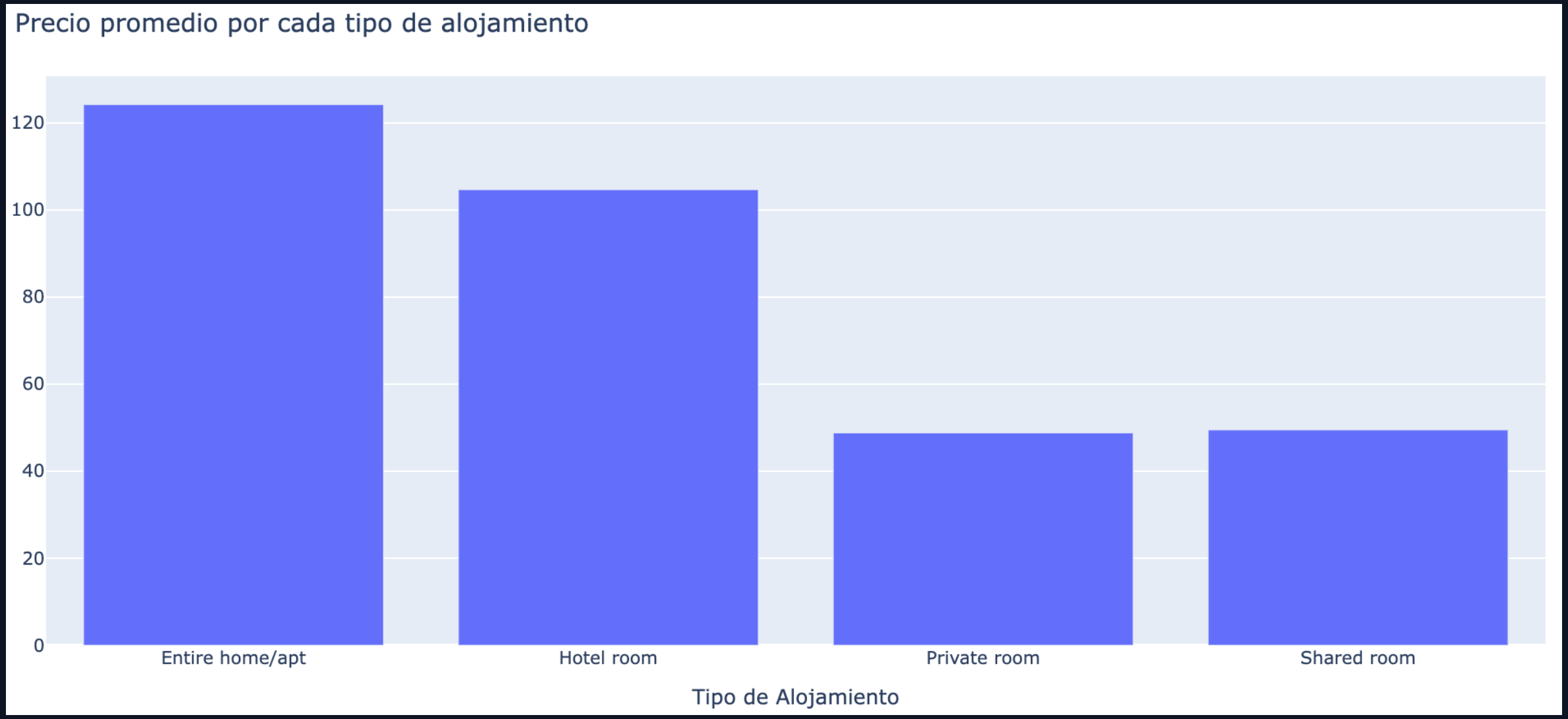
Cada año que pasa, las personas tienen más posibilidades de viajar a otros países y uno de los más atractivos para los países sudamericanos es España, principalmente por el idioma. España se caracteriza por tener ciudades muy turísticas en el verano europeo, como Barcelona, Madrid, la isla de Mallorca, entre otras. Existen muchas aplicaciones para alojar en distintas ciudades del mundo, dentro de las cuales su precio depende de muchos factores, como cercanía a atractivos turísticos, los servicios que ofrecen, la capacidad de personas admitidas, etc.

Este proyecto busca generar un modelo que permita estimar el precio de renta de alojamientos turísticos, ya sea de una habitación privada o de la renta de una casa/apartamento en la ciudad de Barcelona, comprendiendo las relaciones entre las características de los alojamientos y sus respectivas rentas, con el fin de ayudar a los anfitriones a ajustar sus valores de renta y a los que quieran incursionar en esta área de arrendar alojamientos turísticos tomando la mejor decisión.

Datos iniciales



Datos iniciales



Resumen de los datos originales

Data Summary

dataframe	Values
Number of rows	10000
Number of columns	18

Data Types

Column Type	Count
int64	8
float64	6
string	2
bool	2

bool

column_name	true	true rate	hist
is_instant_bookable	5806	0.58	
has_availability	10000	1	

En primera instancia, se seleccionaron los *features* que se creían importantes para el *target*, obteniendo el siguiente resumen de cada componente del modelo:

Dado que 'has_availability' solo está compuesto por valores True, eliminamos esta variable

number

column_name	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
price	254	2.54	100	86	6	45	75	120	500	
latitude	0	0	40	1.4	37	40	40	41	42	
longitude	0	0	0.28	3.2	-6	-3.7	2.2	2.9	4.3	
cant_comodidades	0	0	21	11	1	13	19	27	99	
accommodates	0	0	4.3	2.6	1	2	4	6	29	
bathrooms	74	0.74	1.6	0.99	0	1	1	2	13	
bedrooms	70	0.7	1.9	1.4	0	1	2	3	50	
beds	45	0.45	2.9	2.3	0	1	2	4	30	
minimum_nights	0	0	5	18	1	1	2	4	1100	
maximum_nights	0	0	760	500	1	62	1100	1100	1100	
availability_30	0	0	13	12	0	0	11	26	30	
availability_60	0	0	29	23	0	2	29	53	60	
availability_90	0	0	46	34	0	10	49	80	90	
availability_365	0	0	190	130	0	66	180	320	360	

string

column_name	NA	NA %	words per row	total words
room_type	0	0	2	20000
city	0	0	1	10000

Dataset acotado

Ya que Barcelona junto con Entire home/apt y Private room son los datos con mayor frecuencia en el dataset, le daremos énfasis a estos para crear un nuevo conjunto de datos, manteniendo todas las columnas anteriormente mostradas, pero eliminando la columna 'city' y encodeando 'room_type':

- 0 o False: Entire home/apt
- 1 o True: Private room

Data Summary

dataframe	Values
Number of rows	2678
Number of columns	16

Data Types

Column Type	Count
int64	8
float64	6
bool	2

number

column_name	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
price	43	1.61	84	76	6	35	60	100	500	
latitude	0	0	41	0.015	41	41	41	41	41	
longitude	0	0	2.2	0.018	2.1	2.2	2.2	2.2	2.2	
cant_comodidades	0	0	19	9.9	1	12	17	24	99	
accommodates	0	0	3.3	2.3	1	2	2	4	16	
bathrooms	66	2.46	1.4	0.64	0	1	1	2	8	
bedrooms	15	0.56	1.6	0.95	1	1	1	2	8	
beds	17	0.63	2.2	1.9	1	1	2	3	20	
minimum_nights	0	0	7	14	1	1	2	3	300	
maximum_nights	0	0	710	510	1	60	1100	1100	1100	
availability_30	0	0	12	11	0	0	9	23	30	
availability_60	0	0	27	23	0	2	24	50	60	
availability_90	0	0	44	34	0	8	46	78	90	
availability_365	0	0	180	140	0	49	180	320	360	

bool

column_name	true	true rate	hist
is_instant_bookable	1309	0.49	
room_type	1442	0.54	

Dataset final

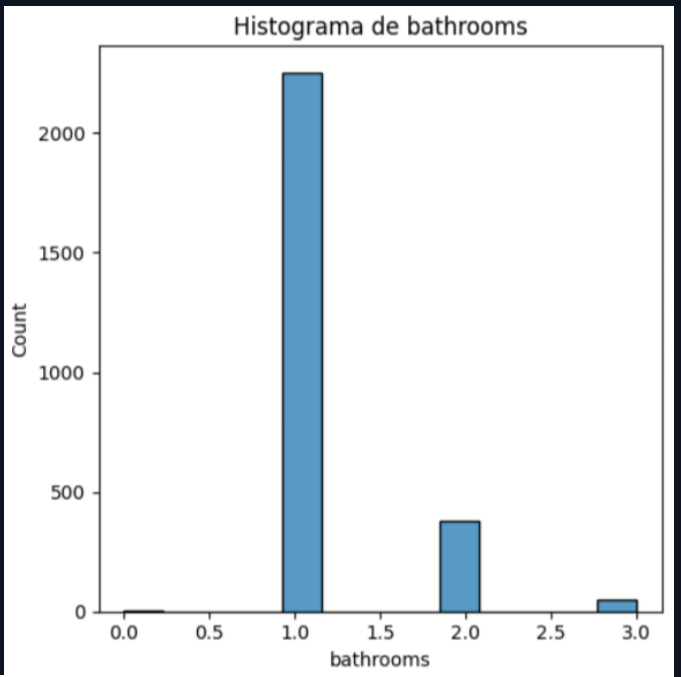
Para identificar los *outliers*, primero se dividió el dataset en 2 conjuntos de datos: Alojamientos que pertenecen a Entire home/apt y Alojamientos que pertenecen a Private room, esto con el motivo de que tienen características muy distintas y un dato común de un tipo de habitación podría ser *outlier* del otro. El método utilizado para identificar estos *outliers* fue IQR, eliminando solo este dato extremo (y así no eliminar la fila completa) y para imputar los valores faltantes se empleó la mediana ya que los datos no tenían una distribución normal, por lo que el promedio no pudo ser utilizado.

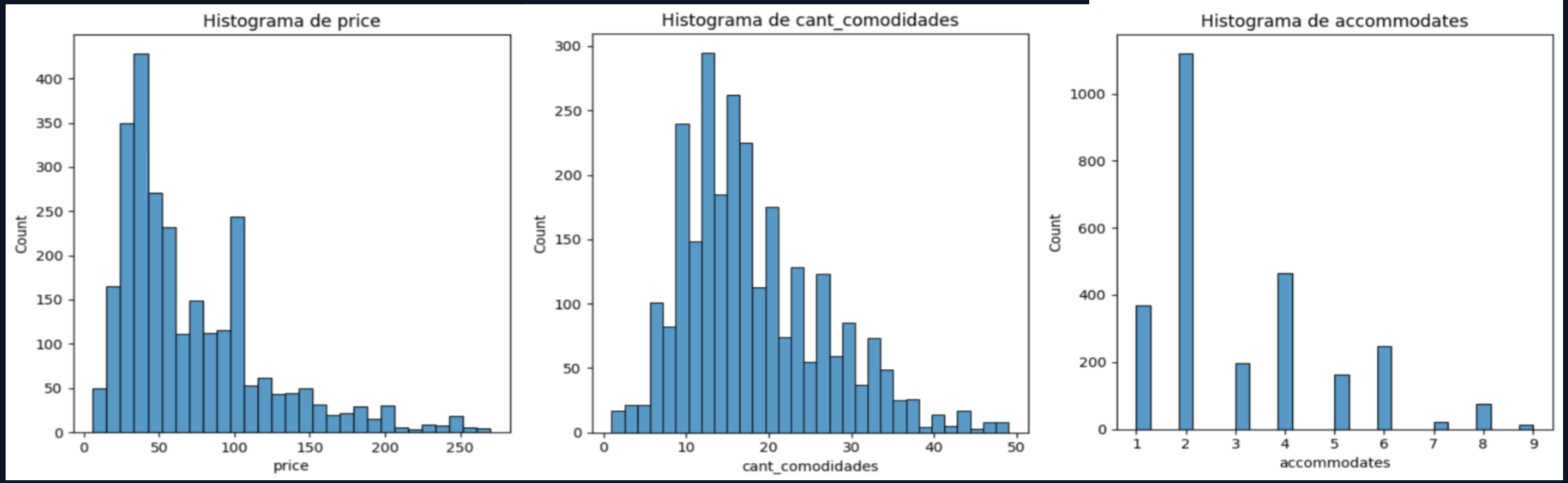
number

column_name	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
price	0	0	71	48	6	35	55	99	270	
latitude	0	0	41	0.015	41	41	41	41	41	
longitude	0	0	2.2	0.018	2.1	2.2	2.2	2.2	2.2	
cant_comodidades	0	0	18	8.6	1	12	16	23	49	
accommodates	0	0	3.1	1.8	1	2	2	4	9	
bathrooms	0	0	1.2	0.42	0	1	1	1	3	
bedrooms	0	0	1.5	0.92	1	1	1	2	6	
beds	0	0	1.9	1.4	1	1	1	3	7	
minimum_nights	0	0	2.2	1.1	1	1	2	3	8	
maximum_nights	0	0	710	510	1	60	1100	1100	1100	
availability_30	0	0	12	11	0	0	9	23	30	
availability_60	0	0	27	23	0	2	24	50	60	
availability_90	0	0	44	34	0	8	46	78	90	
availability_365	0	0	180	140	0	49	180	320	360	

Análisis univariado

A continuación, se muestra el análisis univariado de algunas variables del modelo con el nuevo dataset que corresponde a la ciudad de Barcelona y tipo de habitación Entire home/apt y Private room

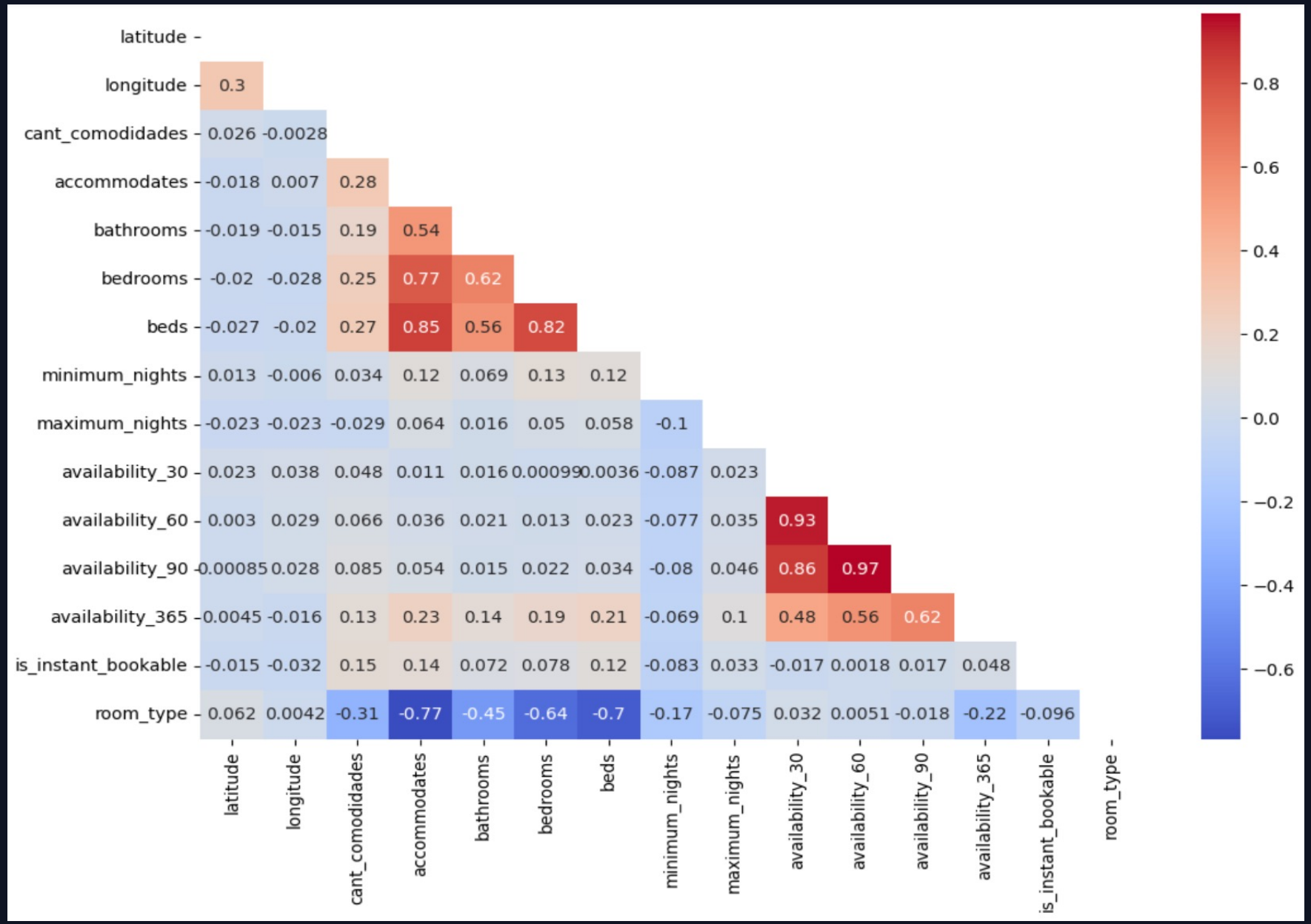




Análisis bivariado

Dejando de lado el *target*, se aprecia una cierta correlación entre algunas variables, por lo que se quitan del modelo (*beds, availability_60,* y *availability_90).*

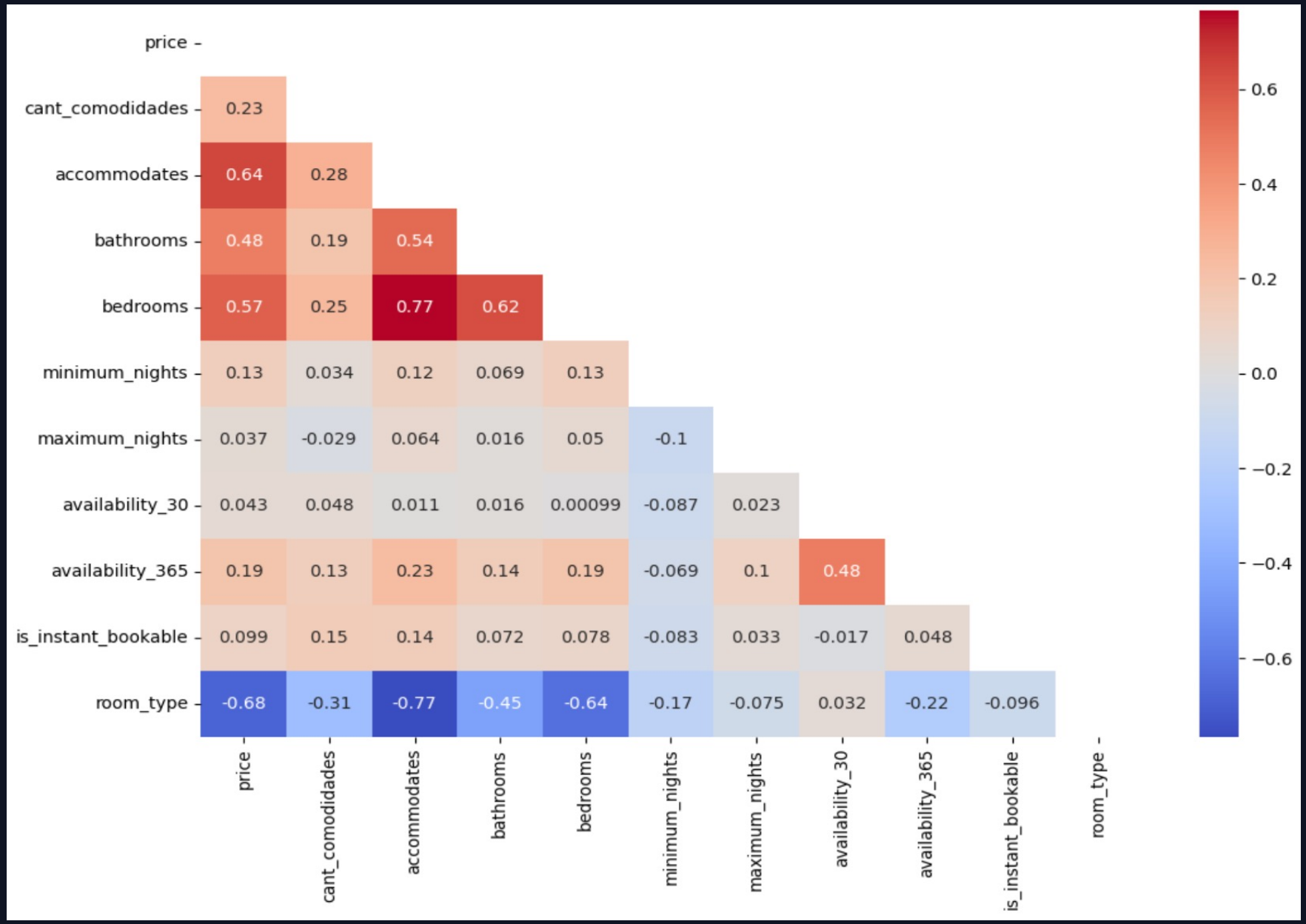
Además, para no generar demasiada variabilidad en el modelo, se eliminan las columnas *longitude* y *latitude*)



Análisis bivariado

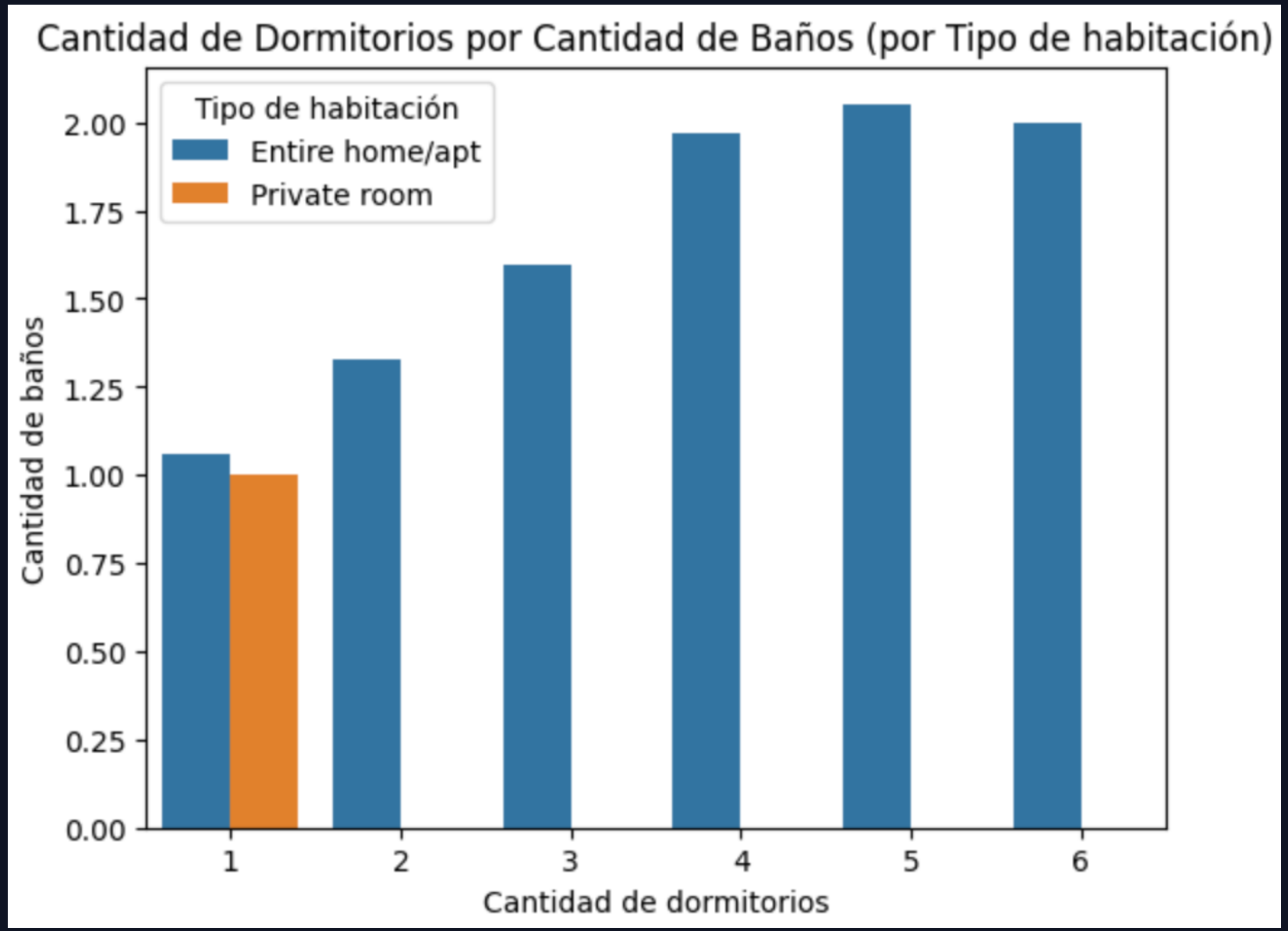
Se realiza un nuevo análisis bivariado, considerando el *target* y los features finales.

En este caso, ya no se presenta una gran correlación entre los *features* y además, el *target price* no está altamente correlacionado a ninguna variable.

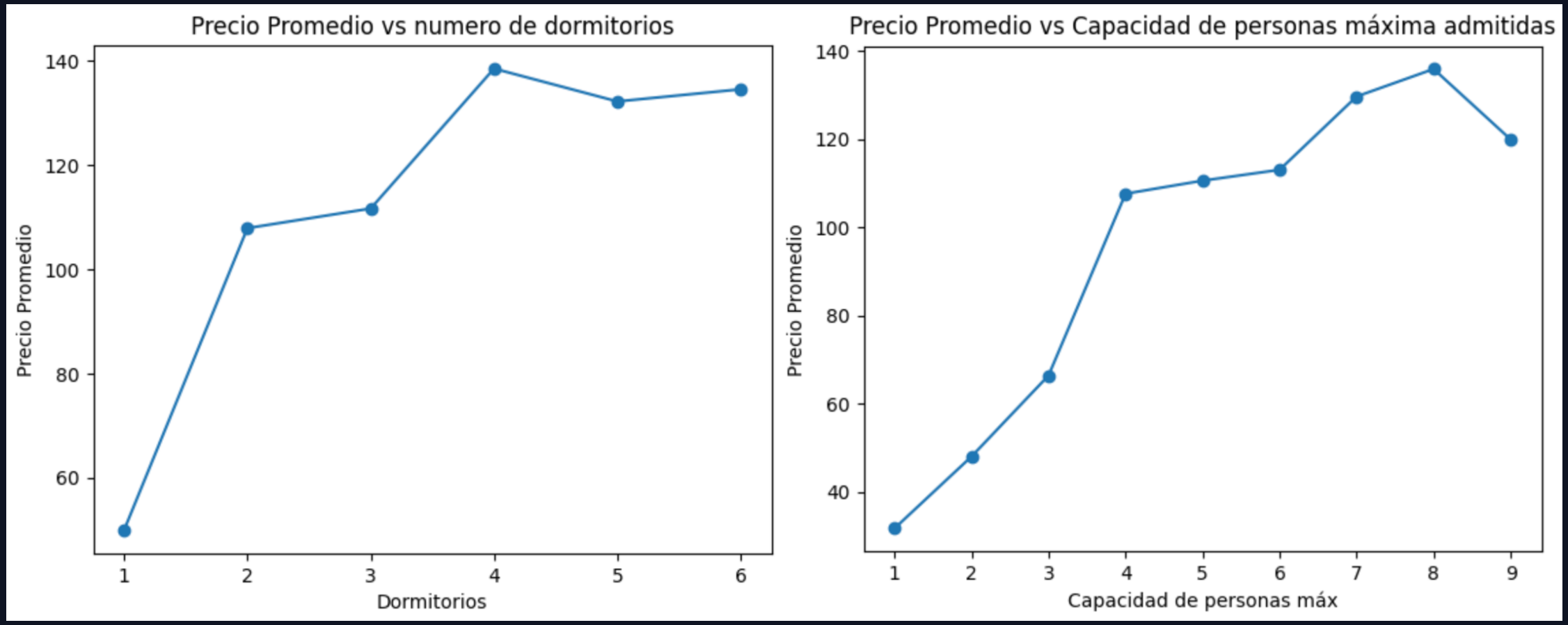


Dormitorios por cantidad de baños

Según vemos, para el caso de *Private room,* como su nombre lo dice, solo posee 1 dormitorio y en promedio ofrece 1 baño. En el caso de *Entire home/apt*, este cuando posee 1 dormitorio, también ofrece un baño, pero a medida que aumenta la cantidad de dormitorios, en promedio, aumenta la cantidad de baños hasta un máximo de 5 dormitorios



Respuestas a las hipótesis de interés



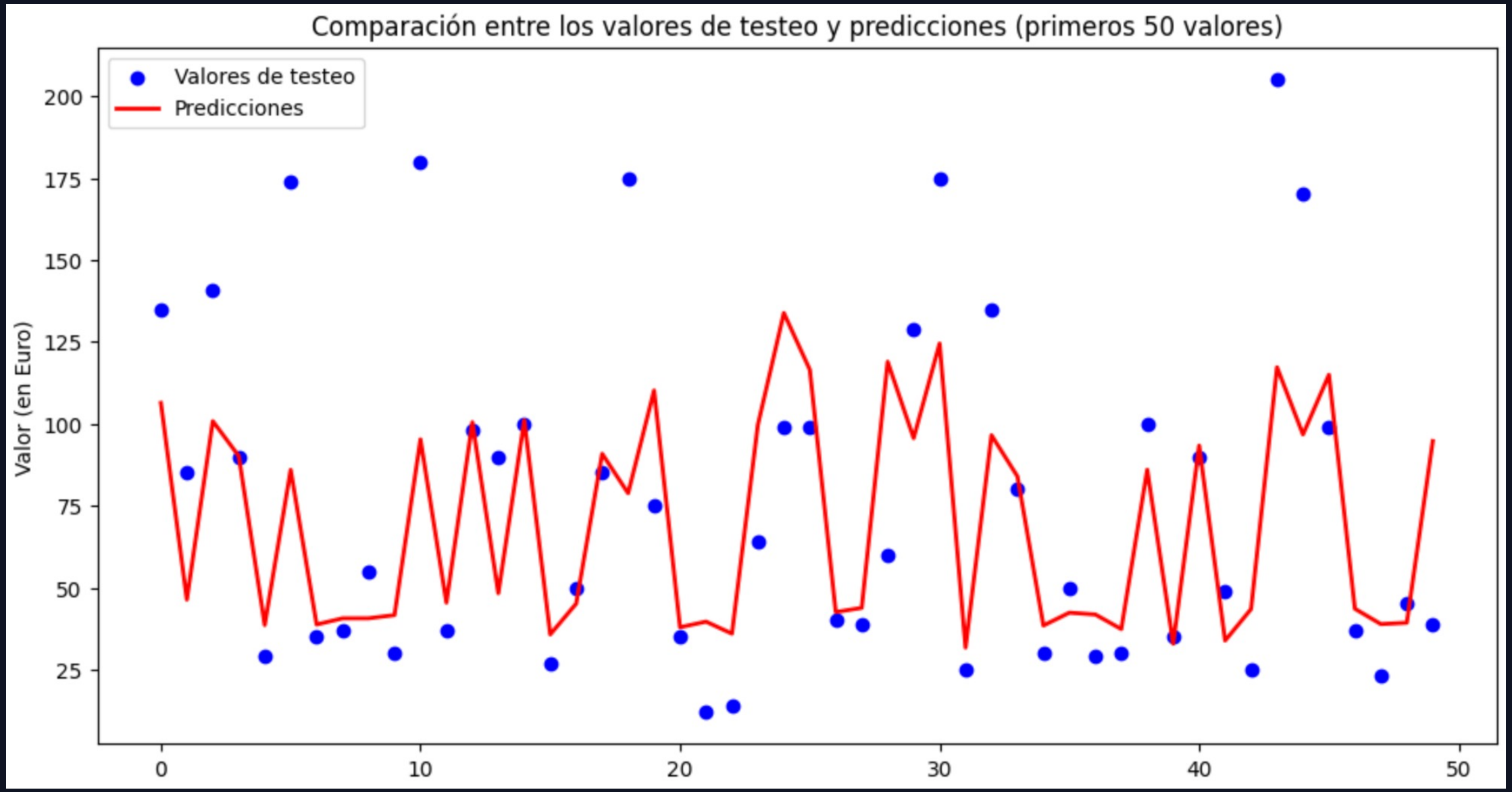
Según vemos en el gráfico de la izquierda, existe una correlación positiva entre el precio promedio de un alojamiento y la cantidad de dormitorios que este ofrece, así como en el gráfico de la derecha, que muestra que también existe una correlación positiva (no tan marcada) entre el precio promedio del alojamiento y la cantidad de personas que permite, es decir, a mayor precio del alojamiento, en general, este tendrá más dormitorios y aceptará más personas.

Modelos de Machine Learning

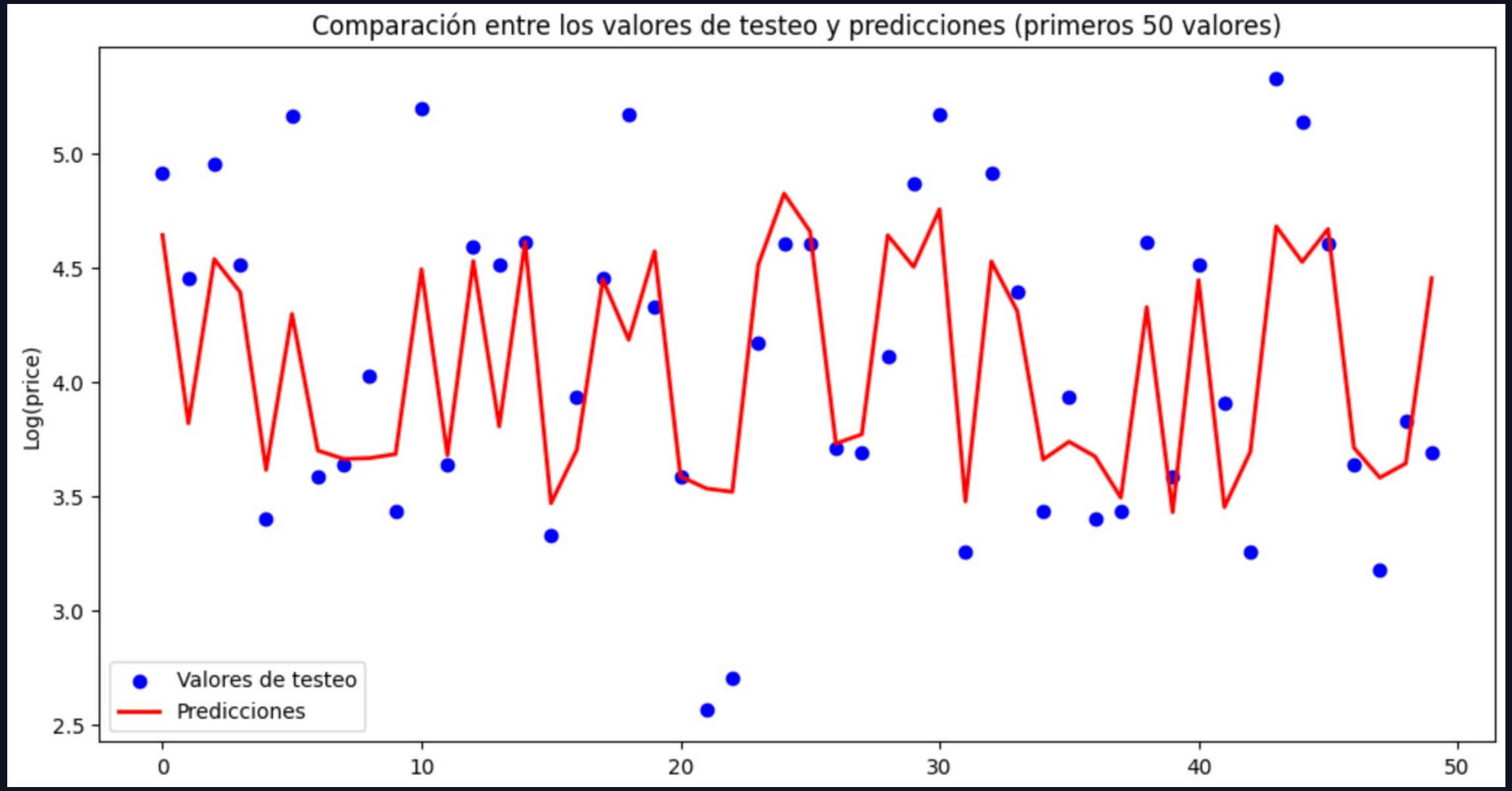
Con el *dataset* final se obtuvieron 2 conjuntos de datos adicionales: transformando todas las variables a su raíz cuadrada y transformando todas sus variables a su logaritmo. Esto con el final de reducir la escala de las variables haciendo así que los datos sean más uniformes.

Con estos 3 *datasets* se realizaron 2 modelos de regresión distintos: Regresión Lineal para el *dataset* transformado tanto en raíz cuadrada como en logaritmo y Regresión de Lasso Lars para el *dataset* con las variables sin transformar

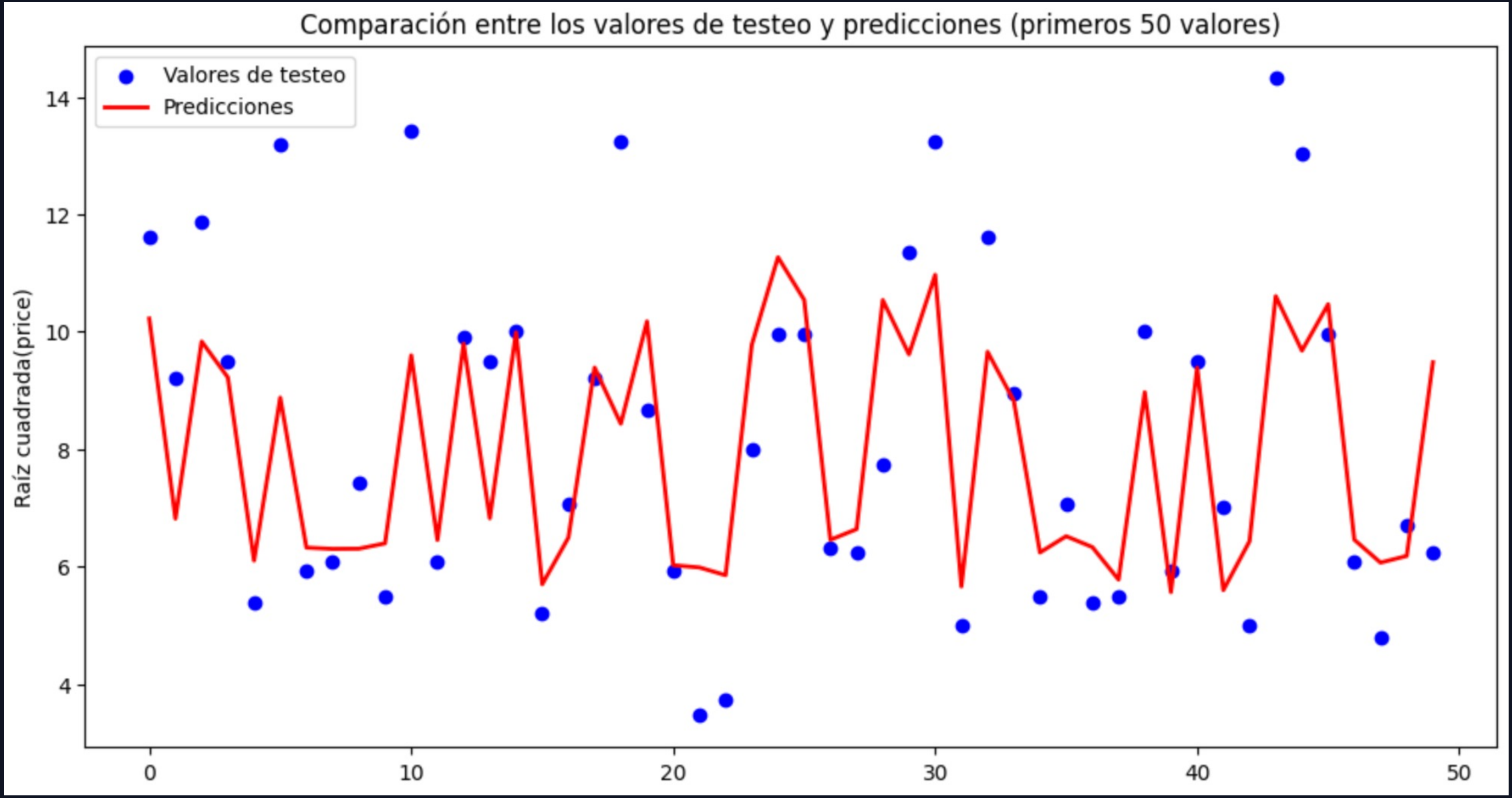
Modelo de Lasso Lars



Modelo de Regresión Lineal (log)

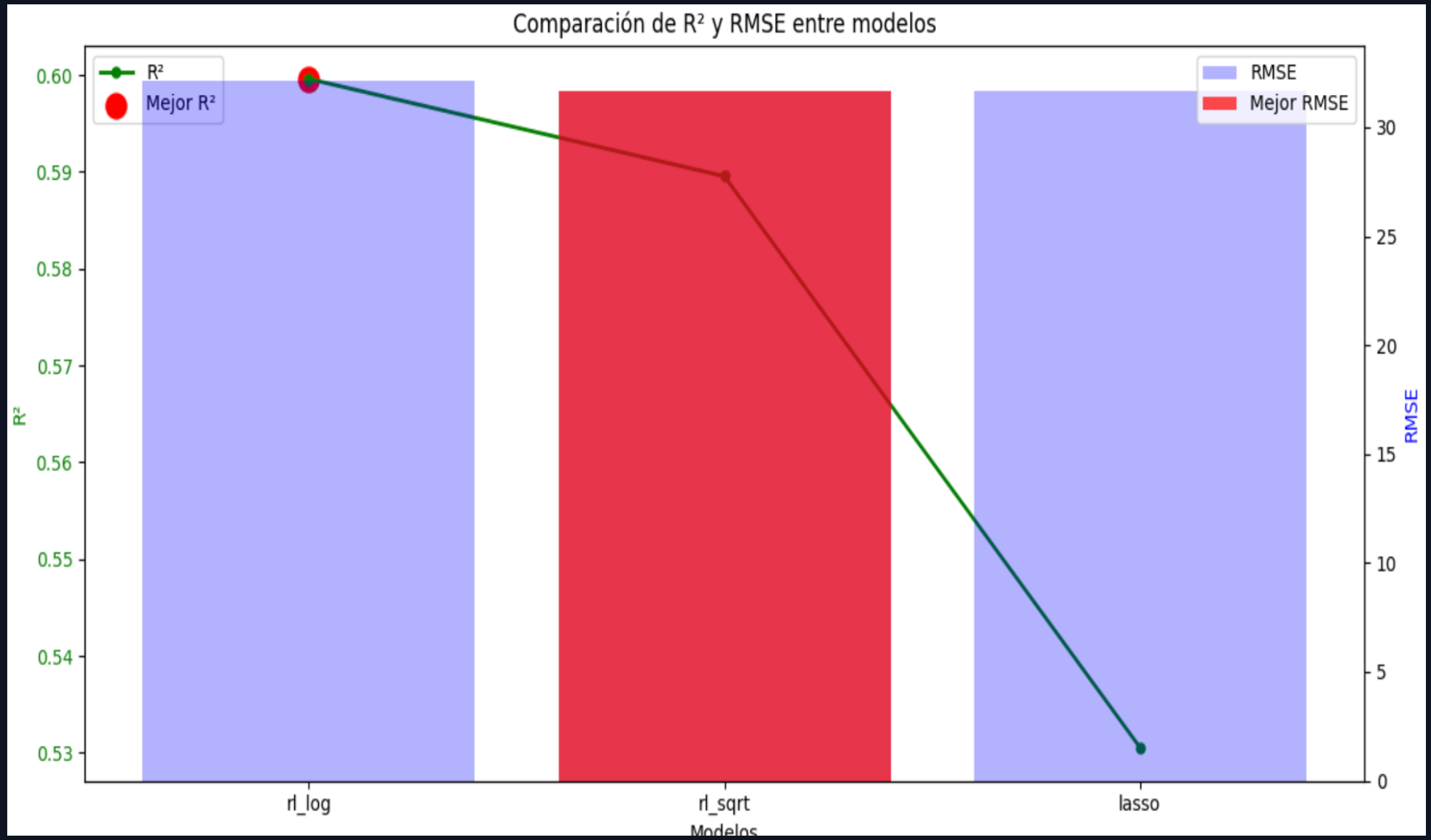


Modelo de Regresión Lineal (sqrt)



Elección de los mejores 2 modelos

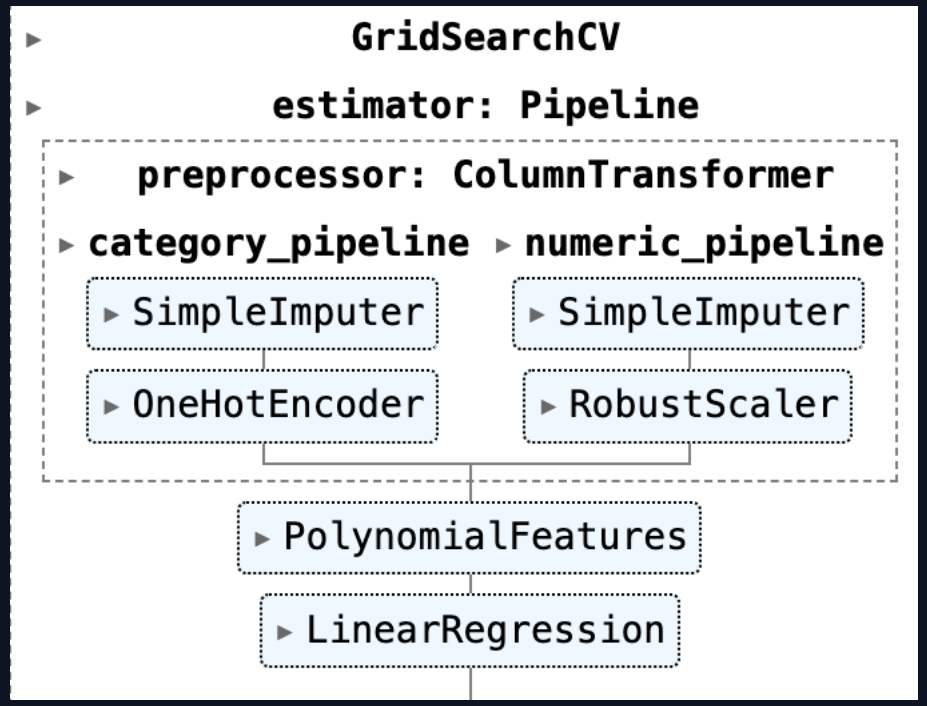
Se realizó una comparación entre los 3 modelos viendo su RMSE y su R^2. De esta manera se concluyó que los mejores 2 modelos fueron ambas regresiones lineales, una por obtener el mejor R^2 (0.6) y la otra por obtener el menor RMSE (31.7)



Mejora de los modelos

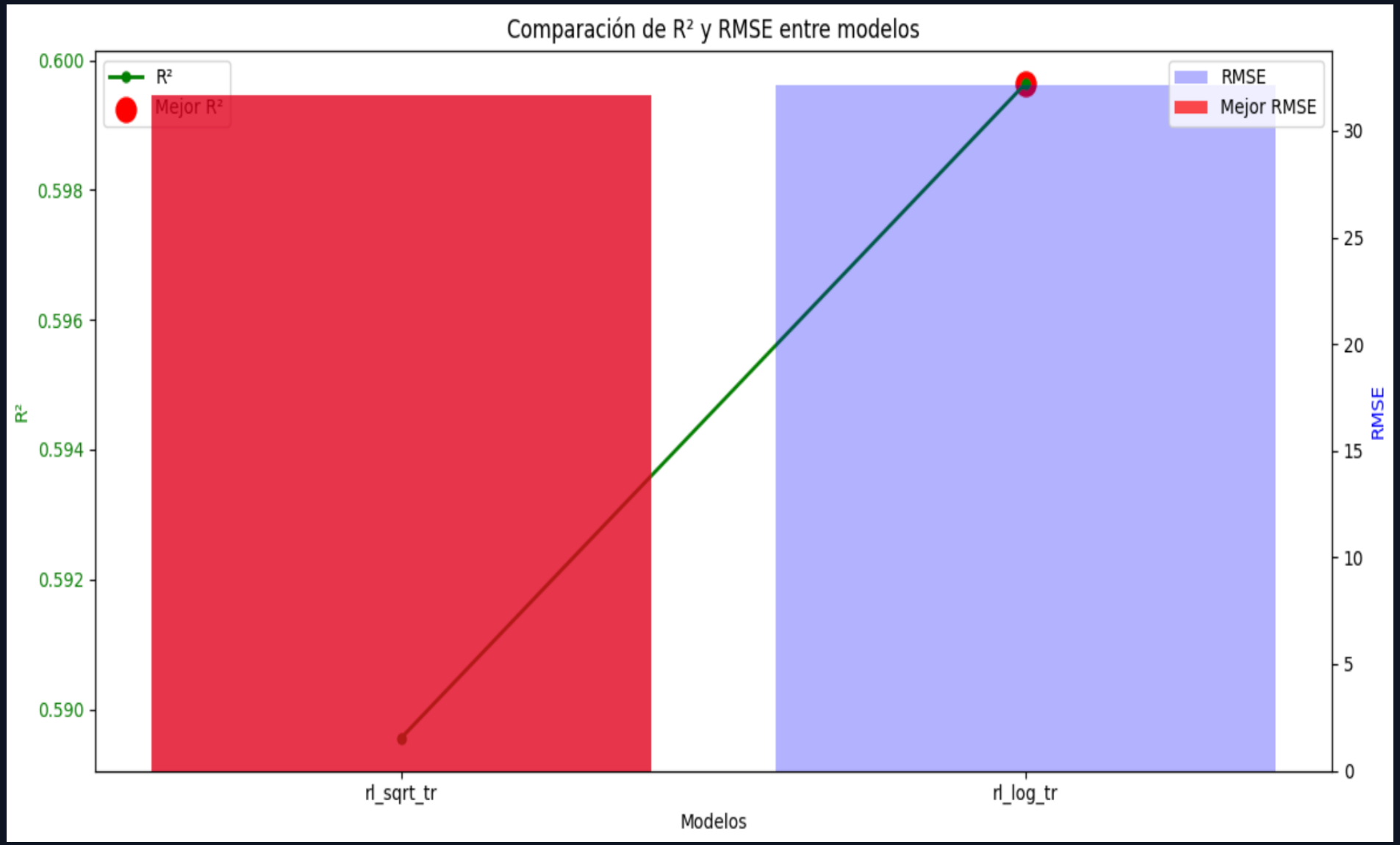
Dado que los mejores 2 modelos correspondían a una Regresión lineal, se realizó un preprocesador de ColumnTransformer para que los modelos sean capaces de encodear e imputar.

Además, se utilizó GridSearch en ambos modelos con el hiperparametro del grado del polinomio entre 1 y 4.



Elección de los mejores 2 modelos

Finalmente, con estas mejoras de ambos modelos, se obtuvieron métricas similares a las anteriores, por lo que el modelo escogido como final fue el que corresponde a las variables transformadas a su raíz cuadrada, ya que si bien no tiene el mejor R^2 (0.59 vs 0.60), tiene el RMSE más bajo (31.7 vs 32.2 euros), y lo que se busca en el modelo, además de un buen valor de R^2, es minimizar el error.



Conclusión

Luego de realizar varios modelos de regresión con 3 tipos de variables (originales, transformadas a raíz cuadrada y a logaritmo), se observó que el modelo con el mejor desempeño fue la Regresión Lineal con la raíz cuadrada de cada variable. Para escogerlo debió compararse con la otra Regresión Lineal, pero puesto que sus valores de R^2 eran similares, se optó por tomar la decisión según el RMSE, los cuales estaban en la misma unidad de medida (Euros), teniendo una diferencia de 0.5 euros.

Tomando en cuenta el gráfico de la comparación de los valores de testeo y las predicciones de este modelo, vemos que las predicciones se ajustan más a los valores de prueba cuando estos son intermedios, es decir, alojamientos que no poseen valores de rentas extremadamente altos ni bajos. Aunque el rendimiento del modelo es aceptable, existe un margen para mejoras, por lo que podría no solo servir como un complemento para tomar mejores decisiones a la hora de poner en renta o querer rentar un alojamiento, si no que podría ser un modelo más robusto, capaz de tomar decisiones optimas según los parámetros requeridos.