UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE CIENCIAS ECONÓMICAS

COMPUTACIÓN CIENTÍFICA ACTUARIAL

Docente a cargo: Del Rosso, Rodrigo

Equipo número 10

Alumnos: Prieto, Augusto D. n° 888.287

Pérez, Tomás A. n° 891.843

Franco, Patricio n° 880.851

Valicenti, Fernando n° 891.634

2020

Introducción

En este trabajo se analizarán los precios de venta de los inmuebles del condado de King (Washington, Estados Unidos), entre los años 2014 y 2015. A partir de éste, se confeccionará un modelo que realice predicciones de precios y se computarán medidas de evaluación para valorar la calidad del mismo.

A su vez, se realizarán pruebas pertinentes del modelo en un conjunto de datos nuevo para el modelo y, de esta manera, poder obtener una estimación del error promedio esperado para nuevas predicciones.

En ánimos de seguir un esquema similar a la metodología de Box-Jenkins[1], el trabajo se encuentra esquematizado de la siguiente manera:

i. Proceso de preparación de los datos
ii. Identificación del modelo
iii. Ajuste del modelo
iv. Validación
v. Evaluación del modelo
vi. Conclusión

Es preciso destacar que, dado el caso de una validación negativa del modelo, se volverá al punto de identificación del modelo. En la última sección se realizarán los comentarios finales acerca de este proyecto, así como también, sugerencias con respecto a la posibilidad de profundizarlo en una ocasión futura.

[1] Box, G. E. P., & Jenkins, G. M. (1973). Some comments on a paper by Chatfield and Prothero and on a review by Kendall. Journal of the Royal Statistical Society. Series A (General), 136(3), 337-352.

i. Proceso de preparación de los datos

Descripción general

El conjunto de datos cuenta con 17 variables, que fueron detalladas de la siguiente manera:

Variable	Descripción
date	Fecha en que se vendió el inmueble.
price	Precio al que se encuentra listado el inmueble.
bedrooms	Cantidad de dormitorios.
bathrooms	Cantidad de baños.
sqft_living	Medidas del living, en pies cuadrados.
sqft_lot	Medidas del lote, en pies cuadrados.
floors	Cantidad de pisos del inmueble.
waterfront	Determina si el inmueble tiene vista a un frente de agua (1= si,0= no).
view	Determina si el inmueble fue visto por algún potencial comprador (1= si,0= no).
condition	Determina la condición del inmueble en una escala del 1 al 5.
grade	Determina la calificación asignada al inmueble, de acuerdo a una escala definida por el condado de King del 1 al 11.
sqft_above	Determina el número de pies cuadrados que ocupa el inmueble, menos los correspondientes al sótano.
sqft_basement	Determina el número de pies cuadrados que ocupa el sótano.
yr_built	Año en el que fue construido.
yr_renovated	Año en el que fue renovado.
sqft_living15	Medidas del living en el año 2015 (implica alguna renovación).
sqft_lot15	Medidas del lote en el año 2015 (implica alguna renovación).

Análisis exploratorio de datos.

En principio, el set de datos contiene 21613 observaciones y no existen registros faltantes. Sin embargo, hay cierta cantidad de datos que carecen de lógica, o bien, que no se corresponde con el dominio de alguna variable explicativa. A continuación, será detallado:

View. Esta variable, se supone, es binaria (0-1) y determina si la casa fue vista por un potencial comprador. Tras inspeccionar los datos, se observa que la variable no es binaria y cuyo dominio es {0, 1, 2, 3 ,4}[2].

Grade. En este caso, según la descripción de la variable, ésta tiene un dominio que incluye a los números enteros entre el 1 y el 11, pero se observan datos con valores 12 y 13. Tras cruzar información, se concluye que el dominio incluye a estos últimos[3].

Bathrooms. El número de baños contiene racionales. Esto es así porque en Estados Unidos se tienen ciertas consideraciones según el equipamiento: Si tiene sólo lavamanos e inodoro se considera "½ baño", si tiene lavamanos, inodoro y ducha se representa como "¾ de baño" y si contiene lavamanos, inodoro, ducha y bañera se considera como "baño completo"[4]. Luego, se puede observar que hay casas en el set de datos con 0 baños. Si bien no son datos lógicos, se dejará como tal por convención del curso.

Bedrooms. Al igual que la variable *Bathrooms*, hay observaciones con 0 habitaciones y serán tratadas de la misma forma, es decir, no habrá modificaciones.

[2] https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=rc. "Total view quality: This is the sum of all view's quality. The view's quality can vary from 0 to 4, in 5 different categories; Puget Sound, City/Territorial, Lake Washington/Sammamish, Mountain, and Small Lake/River."

[3] https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r "Building grade: Represents the construction quality of improvements. Grades run from grade 1 to 13. […]"

[4] https://en.wikipedia.org/wiki/Bathroom#Variations_and_terminology

Ingeniería de atributos

La ingeniería de atributos es una parte muy importante en la preparación de los datos para la posterior elaboración del modelo: a través de la misma se crearán y transformarán variables a partir del conjunto de datos para mejorar la interpretación de los mismos o bien para ajustar los datos a los requerimientos del algoritmo que se utilizará.

Dado que los datos pertenecen a un condado estadounidense, hay algunas variables que refieren a unidades de medida que no son utilizadas con frecuencia en Argentina, por lo tanto, éstas se transformaron con el objetivo de facilitar la lectura de los mismos. Entre ellas se encuentran*: sqft_above, sqft_basement, sqft_lot, sqft_living, sqft_lot15, sqft_living15*. Las mismas están expresadas en pies cuadrados, por lo cual, se crearon nuevas variables con las mismas mediciones expresadas en metros cuadrados: *sqm_above, sqm_basement, sqm_lot, sqm_living, sqm_lot15, sqm_living15*.

Luego, fueron creadas las variables *sqm_total*, que suma los metros cuadrados totales del inmueble incluyendo el sótano, y *home_age*, la cual indica la antigüedad del inmueble desde que se construyó o desde que se remodeló, eligiendo la más reciente entre ambas.

Por último, a partir de la variable *date* se crearon tres nuevas variables –*sale_year, sale_month, sale_day*– que representan el año, mes y día, respectivamente, de venta del inmueble. La creación de estas variables particulares no tiene un objetivo de lectura de datos sino computacional y de funcionalidad del modelo.

Análisis descriptivo

Variables categóricas

Variable	Categoría	Cantidad de observaciones
waterfront	0	21450
	1	163
view	0	19489
	1	332
	2	963
	3	510
	4	319
condition	1	30
	2	172
	3	14031
	4	5679
	5	1701
grade	1	1
	2	0
	3	3
	4	29
	5	242
	6	2038
	7	8981
	8	6068
	9	2615
	10	1134
	11	399
	12	90
	13	13

Variables cuantitativas

Variable	Esperanza	Mediana	Desvío	Mínimo	Máximo	Asimetría	Curtosis
price	540088.14	450000	367127.2	75000	7700000	4.02	34.57
bedrooms	3.37	3	0.93	0	33	1.97	49.05
bathrooms	2.11	2.25	0.77	0	8	0.51	1.28
floors	1.49	1.50	0.54	1	3.50	0.62	-0.49
sqm_living	193.23	177.44	85.33	26.94	1257.90	1.47	5.24
sqm_living15	184.56	170.94	63.67	37.07	576.92	1.11	1.60
sqm_lot	1403.47	707.73	3848.06	48.31	153414.99	13.06	284.98
sqm_lot15	1186.22	707.92	2536.62	60.48	80936.45	9.51	150.71
sqm_above	166.15	144.93	76.93	26.94	874.21	1.45	3.40
sqm_basement	27.08	0	41.12	0	447.79	1.58	2.71
sqm_total	193.23	177.44	85.33	26.94	1257.90	1.47	5.24
home_age	46.61	43.00	28.81	5	120	0.56	-0.53

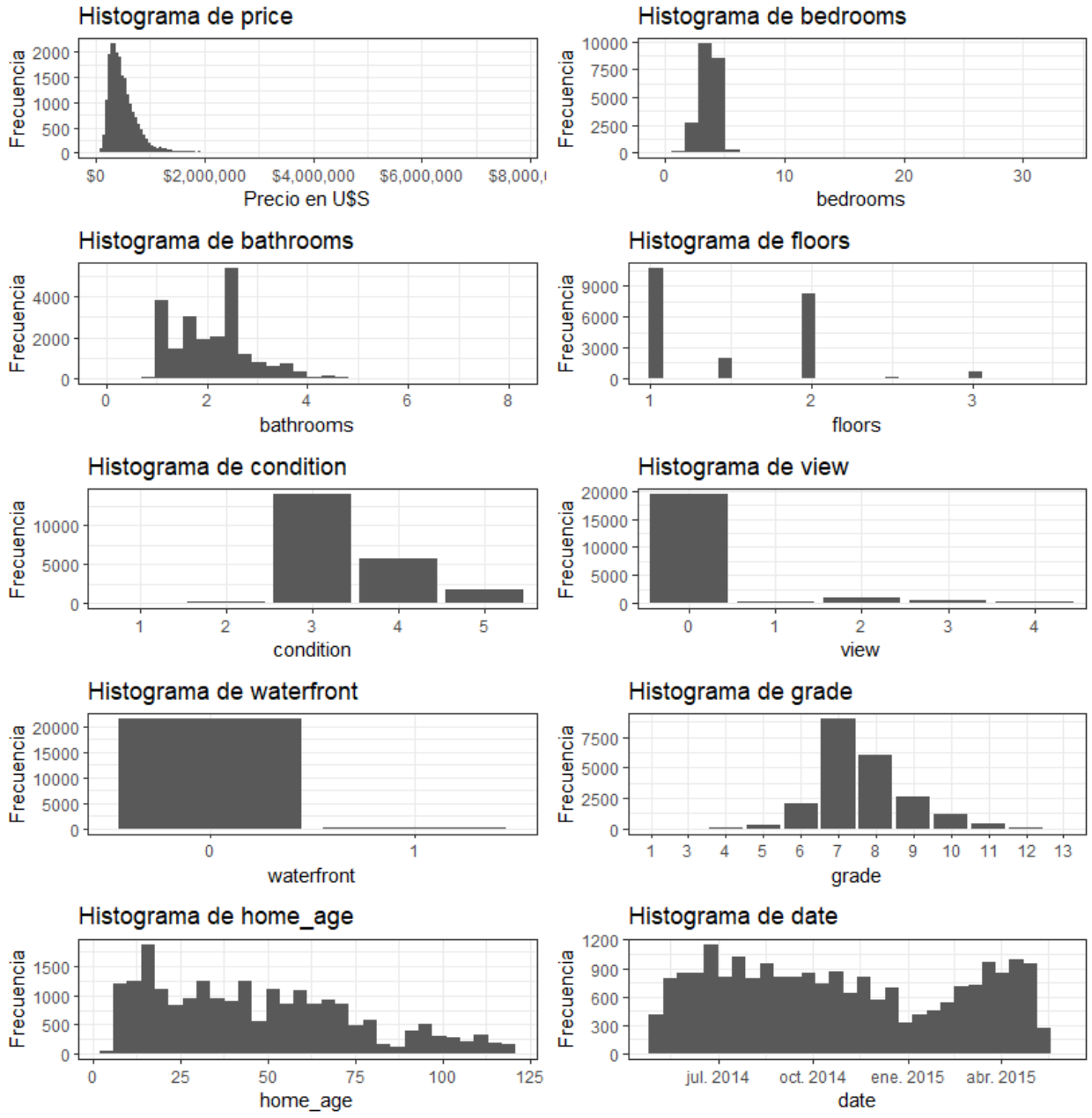
Las medidas descriptivas de las variables cualitativas y cuantitativas serán vistas gráficamente en el siguiente apartado. Sin embargo, es importante calcularlas previamente debido a que facilita la visualización e interpretación de los mismos.

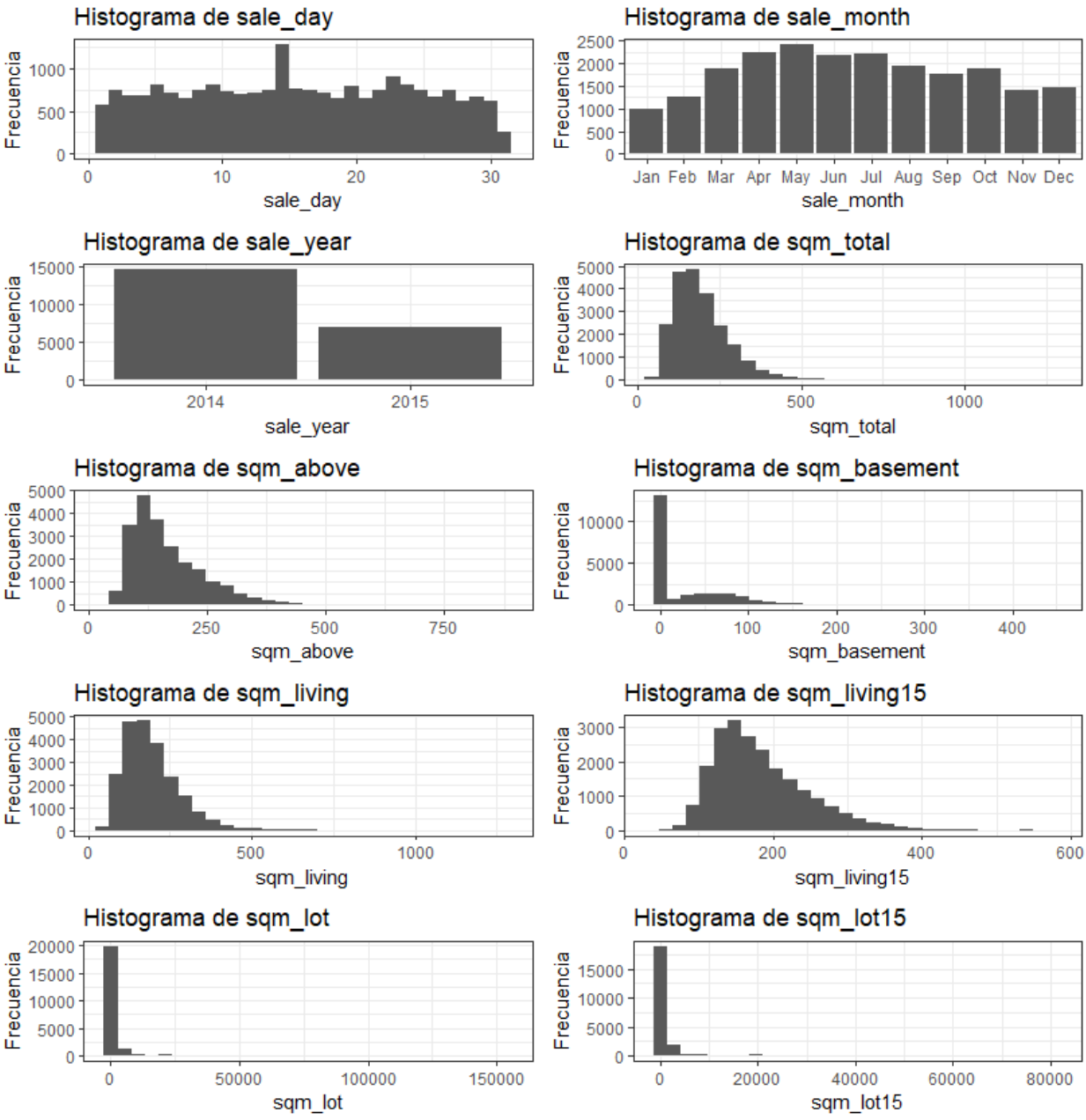
Otro punto de importancia de estas características descriptivas es la posibilidad de ver datos ilógicos o fuera de dominio, que serán luego visualizados en un diagrama de caja, mediante el máximo y mínimo en las variables cuantitativas y a través del conteo de observaciones en las variables cualitativas.

Visualización de datos

Histogramas

Un histograma es una representación gráfica de una variable en forma de barras, donde la superficie de cada barra es proporcional a la frecuencia de los valores representados. Sirven para obtener un panorama de la distribución de variables continuas y conteo de observaciones de variables cualitativas.

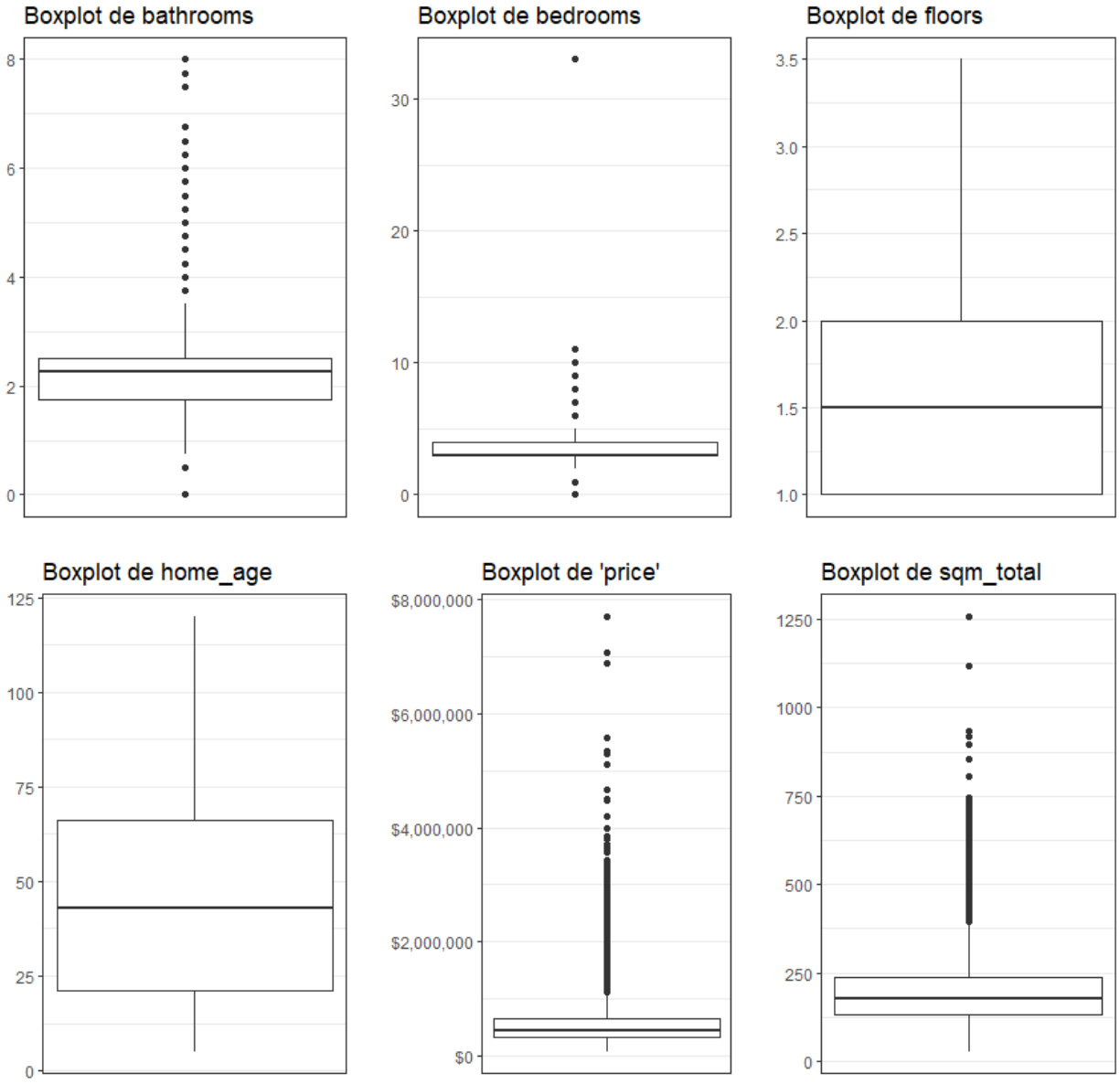


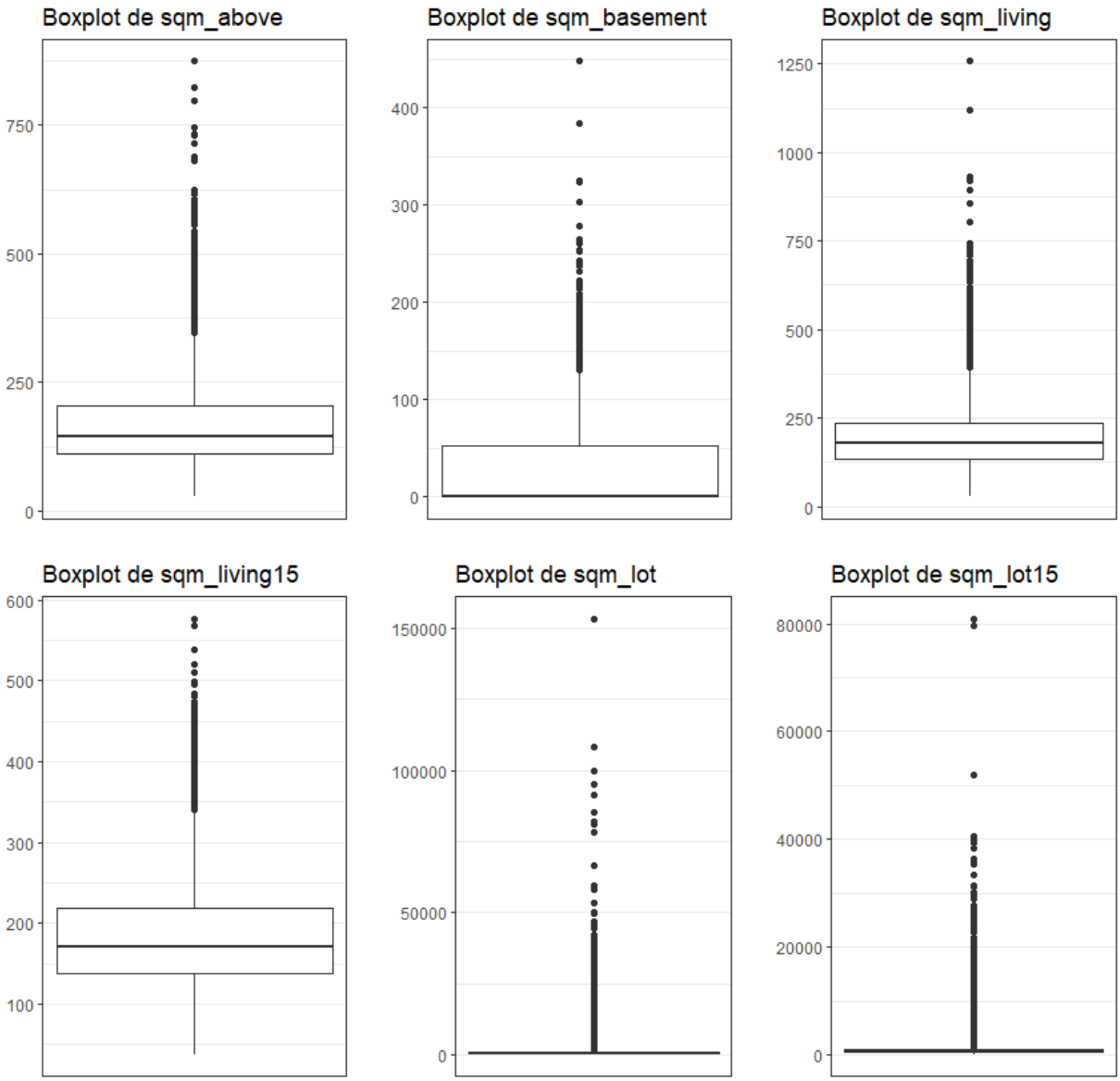


La información que se obtiene de los histogramas es importante no sólo para ver la distribución propia de cada variable cuantitativa, sino para compararla con sus pares y obtener una idea de la posible multicolinealidad asociada. Como se puede ver, este es el caso de algunas de las variables que miden los metros cuadrados del inmueble.

Boxplots o diagrama de caja

El diagrama de caja representa gráficamente los cuartiles de los datos, así como también las observaciones atípicas. Este último punto es de gran importancia para detectar errores en el ingreso de la información, aunque por convención no se realizarán modificaciones de datos que pudieran ser catalogados como erróneos como, por ejemplo, el inmueble con 33 habitaciones que se observa en el boxplot de la variable *bedrooms*.



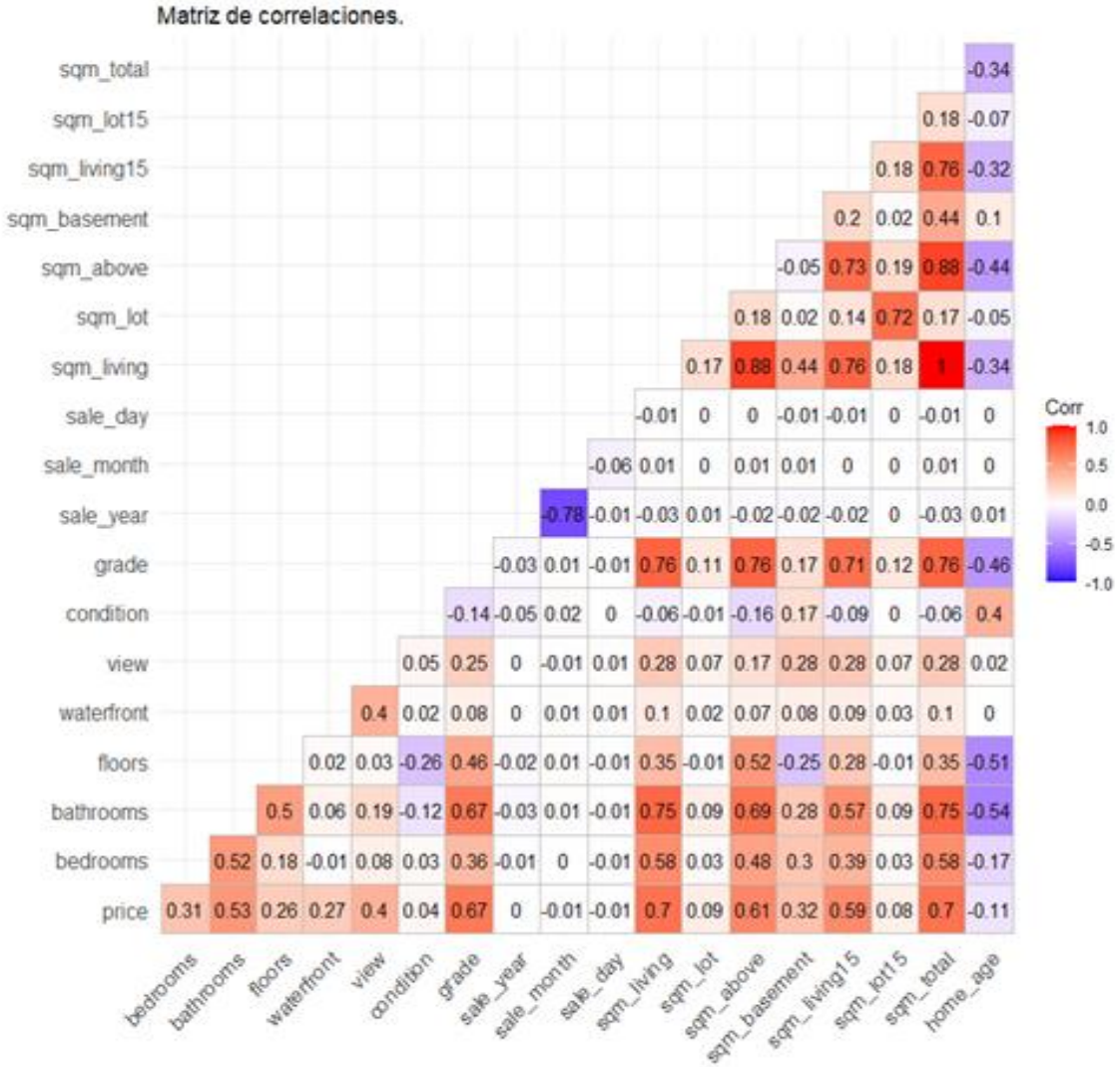


Además de observarse los datos atípicos, que pudieran o no ser un error en el ingreso de información –sobre todo en los casos más extremos–, se destaca también una coincidencia en cuartiles de distribuciones similares antes vistas en los histogramas. Esto eleva, aún más, las sospechas en aquellas variables de medición de metros cuadrados de los inmuebles.

ii. Identificación del modelo

Correlación

La correlación entre variables es importante: ofrece un panorama de elección de variables explicativas para el modelo y, además, de multicolinealidad. Por ejemplo, en la siguiente matriz de correlaciones, se puede observar que la correlación entre *sqm_total* y *sqm_living* es igual a uno. Esto indica que podrían ser linealmente dependientes. Por otra parte, se puede señalar que *sqm_total* está altamente correlacionada con *price*, indicando que los metros cuadrados totales del inmueble podrían explicar su precio.



Selección de variables

Debido al requerimiento de este trabajo, se seleccionaron las diez variables explicativas más importantes, es decir, aquellas que mejor explican a la variable *price*. Este paso, además, ayuda a que no haya un sobre-ajustamiento del modelo, también llamado *overfitting*. Esto significaría una disminución en el *training error*, este es el error que surge de realizar la prueba del modelo con los datos que sirvieron para ajustar el mismo, y a su vez, un aumento en el *testing error*, el cual estima el error de futuras predicciones.

Es importante tener en cuenta el problema de multicolinealidad, es por ello que no se tomó la información recaudada de modelos lineales simples. Es posible que una variable explicativa agregue la misma información que otra variable independiente y cometer el error de agregarla porque lo valida un modelo simple. Luego, cuando ésta conforme el modelo lineal múltiple, aumenta el *testing error.* Tras la sospecha originada en los gráficos, así como también en la matriz de correlación, se calculó el VIF (*Variance Inflation Factor*). Por estar fuera del ámbito de este trabajo, no se entrará en detalle de este concepto: alcanzará con saber que nos indica el grado en que puede ser explicada una variable explicativa por el resto de las variables, exceptuando la variable dependiente. Para este proyecto, se tomó un umbral de tolerancia del VIF igual a 10. Si se superara este número, significa que existe multicolinealidad. Los resultados indican que *sqm_above, sqm_living y yr_built* no agregan información adicional al modelo. Es lógico dado que *sqm_total* fue creada con gran representación de sqm_above, y *home_age* fue creada a partir de *yr_built* (la antigüedad de un inmueble está indicado en gran parte por el año de construcción, excepto por los renovados, como es de esperar).

Luego, se utilizó el algoritmo de *forward selection* para seleccionar las variables que formarán parte del modelo. Este algoritmo comienza con un modelo sin variables explicativas y luego agrega un variable por paso. La variable elegida en cada paso será determinada por una medición –en este caso, AIC– de cada modelo. Es decir, después de ajustar el modelo sin variables explicativas, se ajusta el modelo lineal simple, tomándose el modelo lineal simple con menor AIC y, luego, se computarán sólo los modelos múltiples con dos variables independientes que incluyan la variable explicativa seleccionada del modelo

lineal simple. Se elige el modelo con menor AIC y así sigue el proceso hasta calcular el modelo lineal múltiple con todas las variables o hasta llegar a un límite de variables dentro del modelo, seleccionado previamente. Este método, probablemente, no será el más correcto, pero si el más eficiente computacionalmente: es lo suficientemente adecuado para explicar a la variable independiente sin tener que calcular todos los modelos posibles (*best subset selection*).

iii. Ajuste del modelo

El modelo para predecir la variable *price*, según lo solicitado, es un modelo lineal múltiple con diez variables explicativas. En este caso, las variables elegidas son: *sqm_total, grade, home_age, yr_renovated, bathrooms, bedrooms, view, sqm_lot15* y *condition*. A continuación, se detallará el resultado obtenido tras aplicar el algoritmo de *forward selection*:

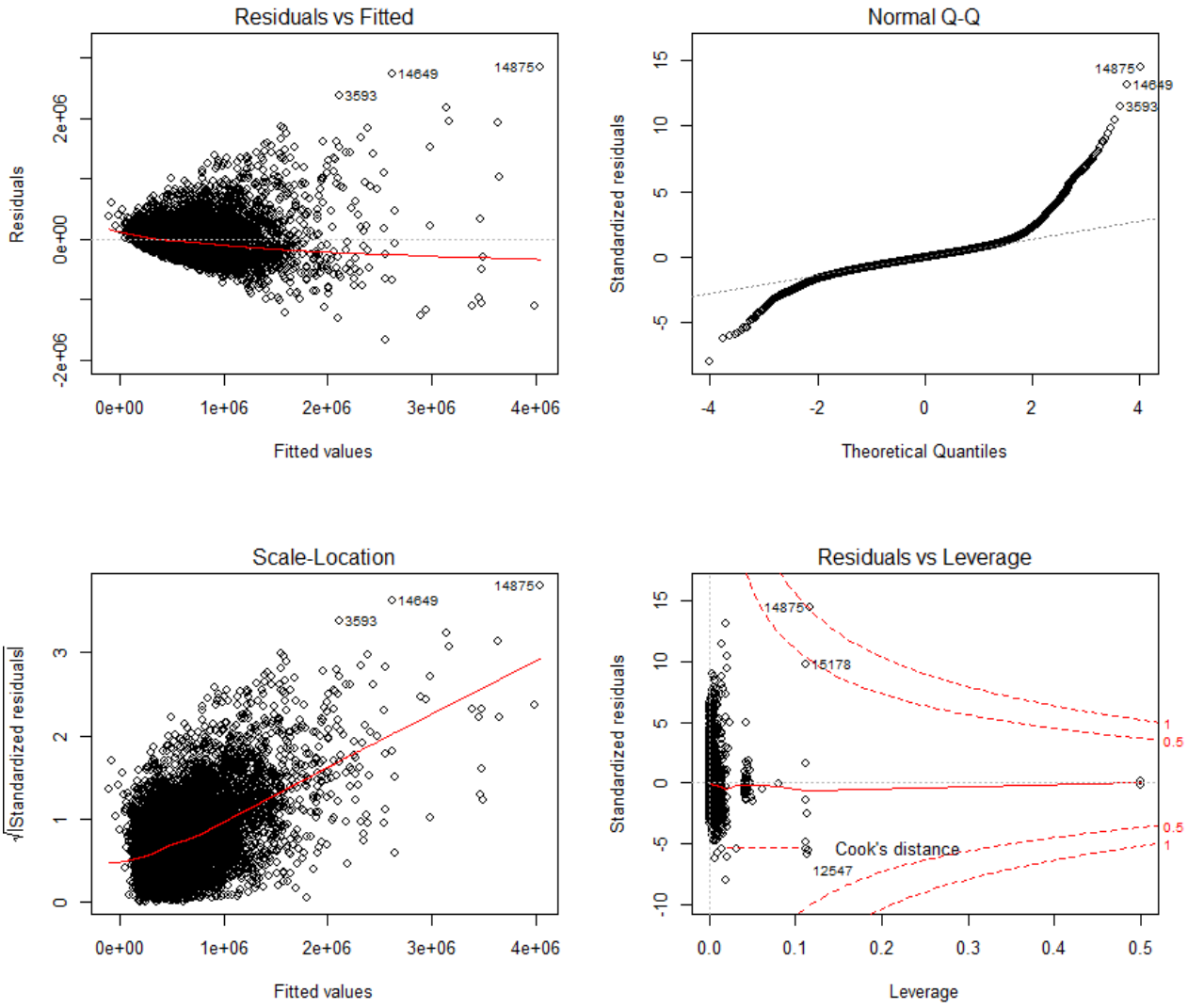
$$\begin{aligned} price = & \beta_0 + \beta_1 \cdot sqm\_total + \beta_2 \cdot grade3 + \beta_3 \cdot grade4 + \beta_4 \cdot grade5 + \beta_5 \cdot grade6 + \beta_6 \cdot grade7 + \\ & \beta_7 \cdot grade8 + \beta_8 \cdot grade9 + \beta_9 \cdot grade10 + \beta_{10} \cdot grade11 + \beta_{11} \cdot grade12 + \beta_{12} \cdot grade13 + \\ & \beta_{13} \cdot home\_age + \beta_{14} \cdot yr\_renovated + \beta_{15} \cdot bathrooms + \beta_{16} \cdot bedrooms + \beta_{17} \cdot view1 + \\ & \beta_{18} \cdot view2 + \beta_{19} \cdot view3 + \beta_{20} \cdot view4 + \beta_{21} \cdot waterfront1 + \beta_{22} \cdot sqm\_lot15 + \beta_{23} \cdot condition2 + \\ & \beta_{24} \cdot condition3 + \beta_{25} \cdot condition4 + \beta_{26} \cdot condition5 + \varepsilon \end{aligned}$$

Las variables categóricas son codificadas automáticamente como variables binarias –*dummy variables*- que indican si una observación pertenece o no a dicha categoría (1 y 0, respectivamente). Es por ello que se observan más parámetros que los anticipados.

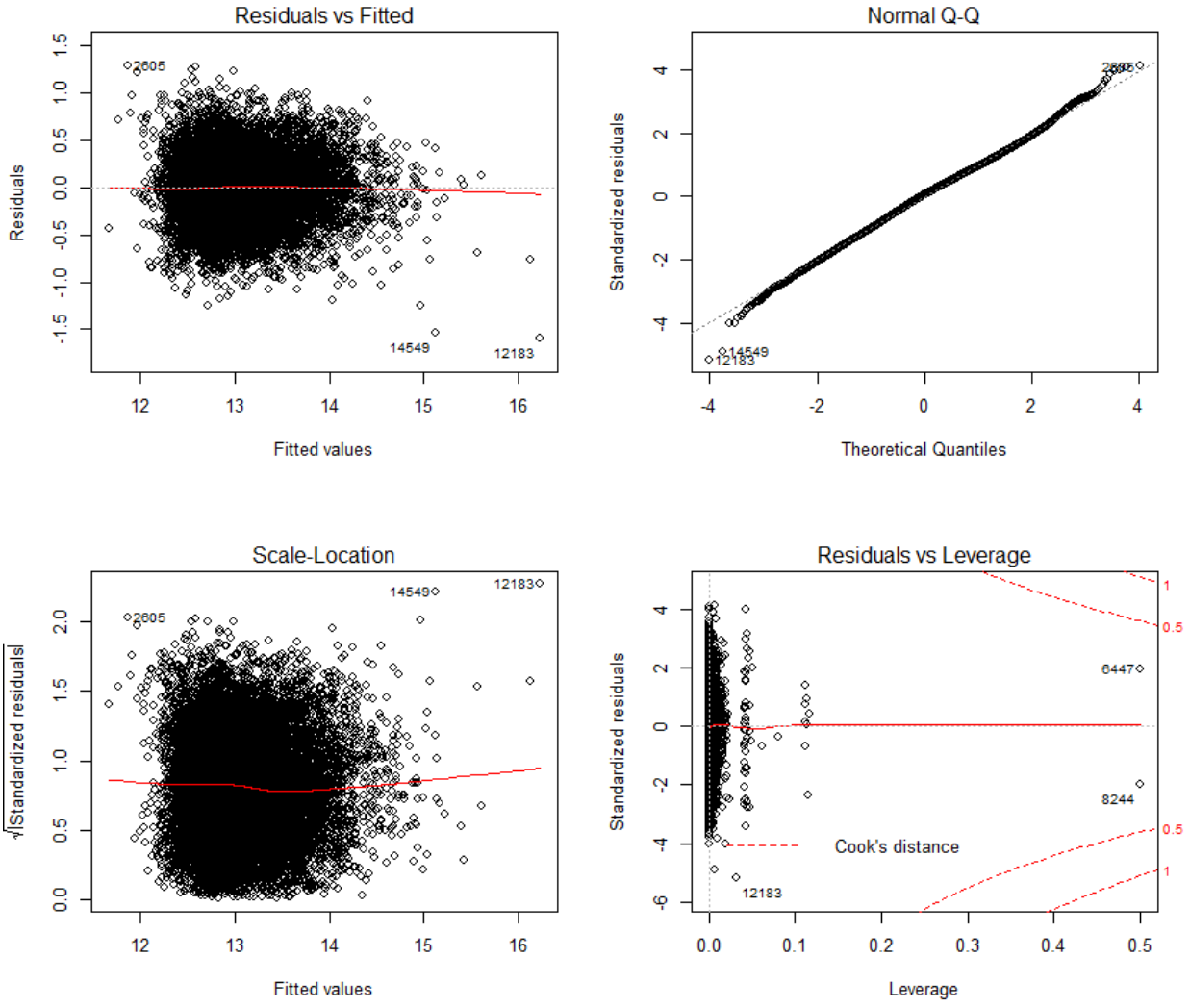
Previo a ajustar el modelo, se realizó una partición aleatoria de datos: el 80% de los mismos se utilizó para ajustar el modelo y el 20% restante, para evaluarlo. Esto es así ya que no es de interés el error que se obtenga con las observaciones que fueron utilizadas para ajustar el modelo, sino el error en nuevas observaciones. Este tema será profundizado más adelante.

iv. Validación del modelo

Tras ajustar el modelo, lo siguiente es verificar las suposiciones que requiere el mismo. Entre ellas se encuentran: homocedasticidad, incorrelación y normalidad de los errores. Antes de realizar las pruebas correspondientes, se puede observar en el gráfico que éstas no se cumplen. Por ejemplo, en el primer gráfico se observa una varianza creciente.



Para reducir la varianza de los residuos, se realizó una transformación logarítmica de la variable dependiente. Los gráficos son más alentadores: no aparenta haber varianza creciente y el gráfico de cuantiles teóricos y muestrales *QQ-plot* es más similar a una distribución normal. Nuevamente, los puntos con influencia en el ajuste del modelo no serán modificados por convención.



Luego, se realizaron las pruebas correspondientes, que arrojaron los siguientes resultados:

Hipótesis nula	P-valor	Decisión
Los errores se encuentran distribuidos normalmente	0.000000e+00	Se rechaza H0
No hay auto-correlación de errores	1.643479e-02	Se rechaza H0
Homocedasticidad	9.978452e-58	Se rechaza H0

Según las pruebas estadísticas, no se cumplen los supuestos del modelo. Esto se puede deber a distintas razones:

- Variables explicativas que no entraron en el modelo, ya sea que pertenezcan al conjunto de datos o variables que no han sido medidas.
- Los puntos de influencia no fueron tratados.
- La relación no es lineal.

Para el propósito de este proyecto, se elegirá este último modelo a pesar de estas dificultades. Quedará para futuros trabajos la puesta a prueba de un modelo no lineal o el tratamiento de los puntos de influencia.

Detalle del modelo:

$$\begin{aligned} log(price) = &\beta_0 + \beta_1 \cdot sqm\_total + \beta_2 \cdot grade3 + \beta_3 \cdot grade4 + \beta_4 \cdot grade5 + \beta_5 \cdot grade6 + \beta_6 \cdot grade7 + \\ &\beta_7 \cdot grade8 + \beta_8 \cdot grade9 + \beta_9 \cdot grade10 + \beta_{10} \cdot grade11 + \beta_{11} \cdot grade12 + \beta_{12} \cdot grade13 + \\ &\beta_{13} \cdot home\_age + \beta_{14} \cdot yr\_renovated + \beta_{15} \cdot bathrooms + \beta_{16} \cdot bedrooms + \beta_{17} \cdot view1 + \\ &\beta_{18} \cdot view2 + \beta_{19} \cdot view3 + \beta_{20} \cdot view4 + \beta_{21} \cdot waterfront1 + \beta_{22} \cdot sqm\_lot15 + \\ &\beta_{23} \cdot condition2 + \beta_{24} \cdot condition3 + \beta_{25} \cdot condition4 + \beta_{26} \cdot condition5 + \varepsilon \end{aligned}$$

A continuación, los resultados de las estimaciones y su correspondiente p-valor:

Coeficiente	Estimación	Error std.	t valor	Pr(>|t|)
(Intercept)	1.151e+01	3.128e-01	36.786	< 2e-16
sqm_total	1.969e-03	5.970e-05	32.972	< 2e-16
grade3	-3.591e-01	3.885e-01	-0.924	0.355241
grade4	-3.431e-02	3.252e-01	-0.106	0.915970
grade5	1.518e-02	3.198e-01	0.047	0.962147
grade6	2.150e-01	3.195e-01	0.673	0.500974
grade7	4.971e-01	3.195e-01	1.556	0.119751
grade8	7.514e-01	3.195e-01	2.352	0.018700
grade9	9.911e-01	3.196e-01	3.101	0.001934
grade10	1.158e+00	3.198e-01	3.621	0.000294
grade11	1.266e+00	3.203e-01	3.953	7.74e-05
grade12	1.334e+00	3.222e-01	4.140	3.49e-05
grade13	1.320e+00	3.374e-01	3.911	9.21e-05
home_age	5.454e-03	1.160e-04	46.996	< 2e-16
yr_renovated	1.534e-04	6.080e-06	25.238	< 2e-16
bathrooms	9.503e-02	5.488e-03	17.317	< 2e-16
bedrooms	-3.552e-02	3.476e-03	-10.220	< 2e-16
view1	1.669e-01	1.953e-02	8.548	< 2e-16
view2	8.791e-02	1.177e-02	7.470	8.42e-14
view3	1.202e-01	1.589e-02	7.566	4.06e-14
view4	2.186e-01	2.452e-02	8.916	< 2e-16
waterfront1	3.236e-01	3.393e-02	9.537	< 2e-16
sqm_lot15	-3.888e-06	9.767e-07	-3.981	6.88e-05
condition2	-4.309e-02	6.951e-02	-0.620	0.535378
condition3	1.470e-01	6.445e-02	2.281	0.022538
condition4	1.543e-01	6.446e-02	2.394	0.016690
condition5	2.189e-01	6.483e-02	3.376	0.000737

v. Evaluación del modelo

En este último apartado, se pondrá a prueba el modelo elegido en el conjunto de datos previamente particionado (20%), para estimar el error esperado en las predicciones de un nuevo conjunto de datos.

La medida de evaluación elegida es conocida como *Mean Absolute Percent Error*, de ahora en adelante, llamada *MAPE.* A continuación, se detalla la forma de calcularlo:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Por sus características, es una de las medidas de precisión más utilizadas, ya que penaliza residuos por exceso y por defecto, pero además se obtiene el porcentaje dividiéndolos por la observación que le corresponde. Tras realizar el cálculo, los resultados obtenidos para el modelo elegido fueron los siguientes:

Training MAPE	Testing MAPE
25.87%	26.25%

La poca diferencia respecto a los *MAPE* de los datos para entrenar el modelo y los datos para probarlo indica que no hay presencia de *overfitting* o sobre-ajustamiento. Además, se espera que, en promedio, el error en nuevas predicciones sea de un 26.25%.

vi. Conclusión

En el expuesto trabajo, se hizo uso de un modelo paramétrico predeterminado, el modelo lineal múltiple, que no cumplió con los supuestos que demanda para la estimación óptima de los parámetros. Si bien no se pudo solucionar este problema, se alcanzó un grado de diferencia entre el training set y el testing set que lo sitúa, por lo menos, como un modelo de referencia de predicción. Lo interesante será, entonces, tratar los puntos de influencia y, dado el caso en el que no lo solucione, ajustar otro tipo de modelo como, por ejemplo, un modelo no paramétrico como bosques aleatorios -*random forest-.*

Bibliografía

- James, Witten, Hastie, Tibshirani (2015). "An Introduction to Statistical Learning, with Applications in R". New York: Springer.
- Edward W. Frees (2010). "Regression Modeling with Actuarial and Financial Applications". New York: Cambridge.
- Hogg, R.V.; McKean, J.W.; and Craig, A.T. (2013), "Introduction to Mathematical Statistics".
- Marko Sarstedt, Erik Mooi (2014). "A concise guide to market research. The Process, Data, and Methods Using IBM SPSS Statistics.". Springer.
- Bhimasankaram Pochiraju, Sridhar Seshadri (2019). "Essentials of business analytics". Springer.