Habi: Prueba Analyst

Andrés Felipe Díaz Rodríguez

Diciembre 2020

1. **Calidad de los datos.** La base de datos proporcionada contiene la información de 14 variables para 2625 inmuebles distintos. La limpieza de los datos empezó por omitir las variables que no tienen relevancia para el análisis propuesto en el ejercicio. Estas son el nombre y el teléfono de contacto de la persona encargada de la venta del inmueble, ya que no proporcionan ninguna información relevante.

   A partir de las 12 variables restantes se revisó la cantidad de observaciones que contenían datos faltantes. La Tabla 1 muestra en la primera columna el nombre de las variables y en la segunda la cantidad de observaciones faltantes para cada variable. Una variable con muchos elementos vacíos es el tipo del inmueble y esto ocurre porque originalmente en la base de Excel tienen tipo "0" -en vez de "Casa" o "Apartamento"- así que se decidió dejarlos en blanco. Sin embargo, esto no es problemático dado que esta variable no tiene casi variabilidad dentro de la muestra, con un 97.66 % de las observaciones marcadas como apartamentos y no se incluirá en el análisis.

   Las demás variables con valores faltantes son el estrato del inmueble, la UPZ en la que está ubicado[1] y su antigüedad. Estas observaciones se deben eliminar de la base de datos analítica ya que estas tres sí son variables clave para el análisis sobre el valor de un inmueble. En total son 483 observaciones que tienen valores faltantes en alguna de estas tres variables, de manera que la muestra final está conformada por 11 variables y 2142 inmuebles que tienen la información completa.

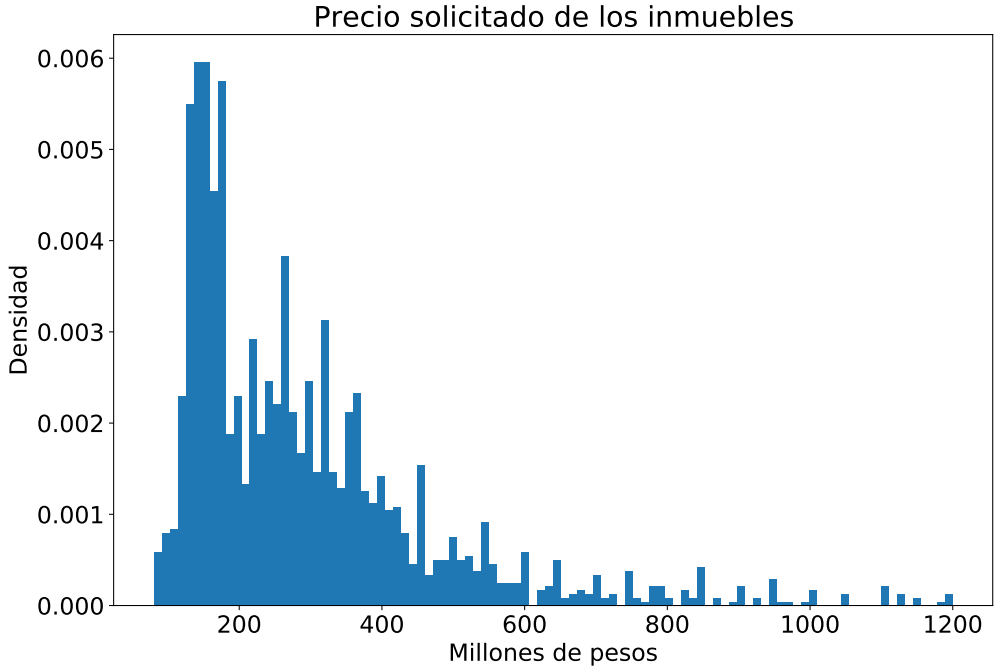
Tabla 1: Cantidad de datos faltantes en las variables de la muestra

Variable	Obs. faltantes
id	0
fuente	0
id_conjunto	0
tipo	204
precio_solicitado	0
area	0
piso	0
garajes	0
ascensores	0
estrato	192
upz	299
antiguedad	114

[1] La gran cantidad de valores faltantes en la UPZ se debe a que se reemplazaron los que estaban marcados como "error" con vacíos.

La variable más importante de la base de datos es el precio de venta solicitado, que está en pesos. La Figura 1 muestra la distribución del precio de los inmuebles de la muestra, cada barra representa la cantidad de inmuebles que tienen un precio correspondiente al eje horizontal. El valor mínimo es de 80 millones y el máximo es de 1.200 millones de pesos. Se observa que hay una gran cantidad de inmuebles alrededor de 150 millones y otro pico alrededor de 300 millones, con una cantidad menor de inmuebles en los precios más altos.

Figura 1: Precio solicitado de los inmuebles



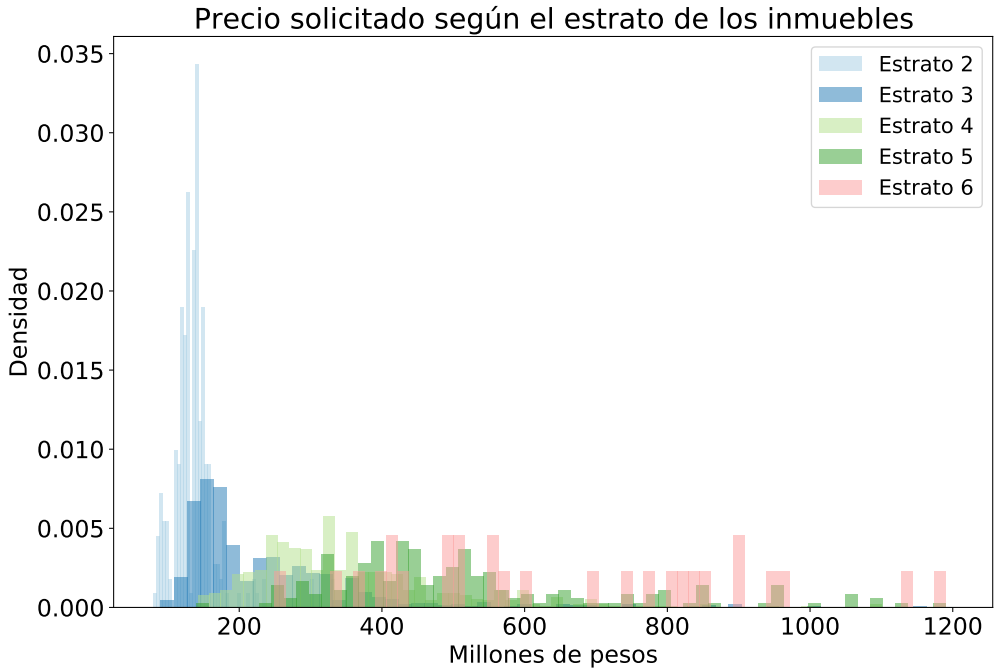
La muestra está dividida entre 1279 inmuebles de estrato 3; 607 de estrato 4; 281 de estrato 2; 31 de estrato 6 y 2 de estrato 1. La Figura 2 divide la distribución de precios según el estrato de los inmuebles y la Tabla 2 muestra el promedio y la desviación estándar de la distribución de precio para cada estrato[2]. Tanto en el histograma como en la tabla con los promedios se observa que el precio de las viviendas va aumentando con el estrato, lo que da cierta confianza sobre la calidad de los datos de la base ya que esto es algo que se espera. Además, se observa que los precios de los inmuebles de estrato 2 son los que están más concentrados, con una desviación estándar mucho menor que los demás estratos, y que debido la baja cantidad de inmuebles de estrato 6 hay una desviación estándar muy alta para este grupo.

Tabla 2: Precios promedio según el estrato de los inmuebles

Estrato	Precio promedio (millones)	D.E.
2	137.885	33.578
3	245.058	150.385
4	354.829	140.053
5	514.961	200.304
6	658.393	247.974

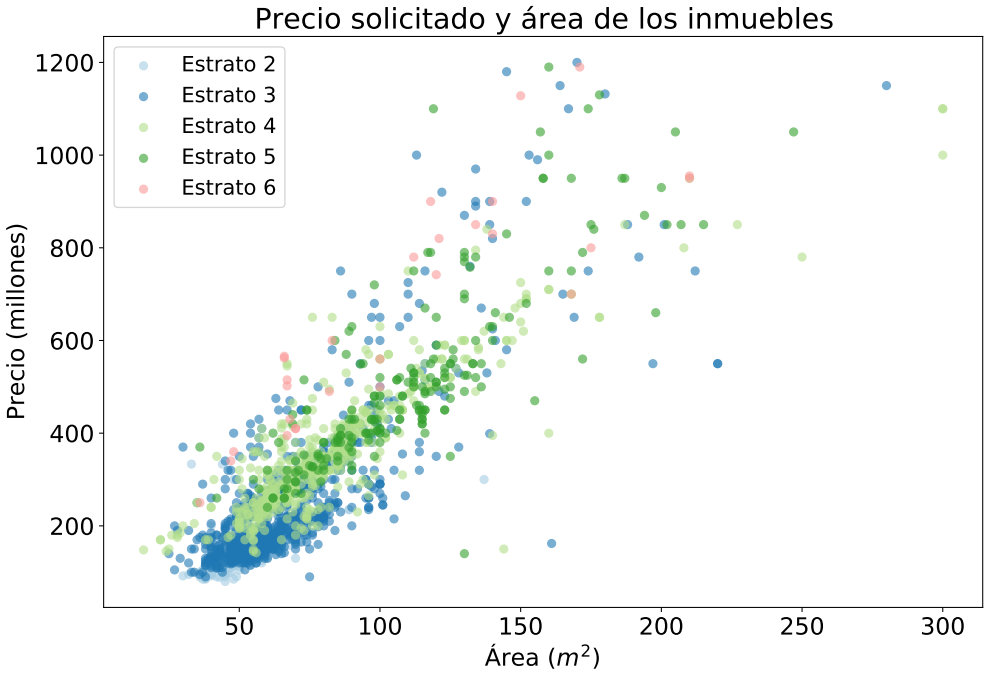
[2]En la gráfica y la tabla se ignoraron los inmuebles de estrato 1 por ser solo dos observaciones.

Figura 2: Precio solicitado de los inmuebles según el estrato



Para poder hacer comparaciones correctas entre los distintos inmuebles de la muestra es necesario hacer el análisis no solo del precio de venta sino de su área. La Figura 3 muestra esa relación: en el eje y está el precio solicitado para la venta y en el eje x está el área en metros cuadrados de los inmuebles y cada punto tiene un color que representa el estrato al que pertenece. Hay una clara relación positiva -entre más extensas son las propiedades, son más costosas- y los inmuebles de estratos 2 y 3 son tienden a ser más pequeños que los de estratos 4, 5 y 6, con algunas excepciones.

Figura 3: Precio solicitado de los inmuebles según el estrato

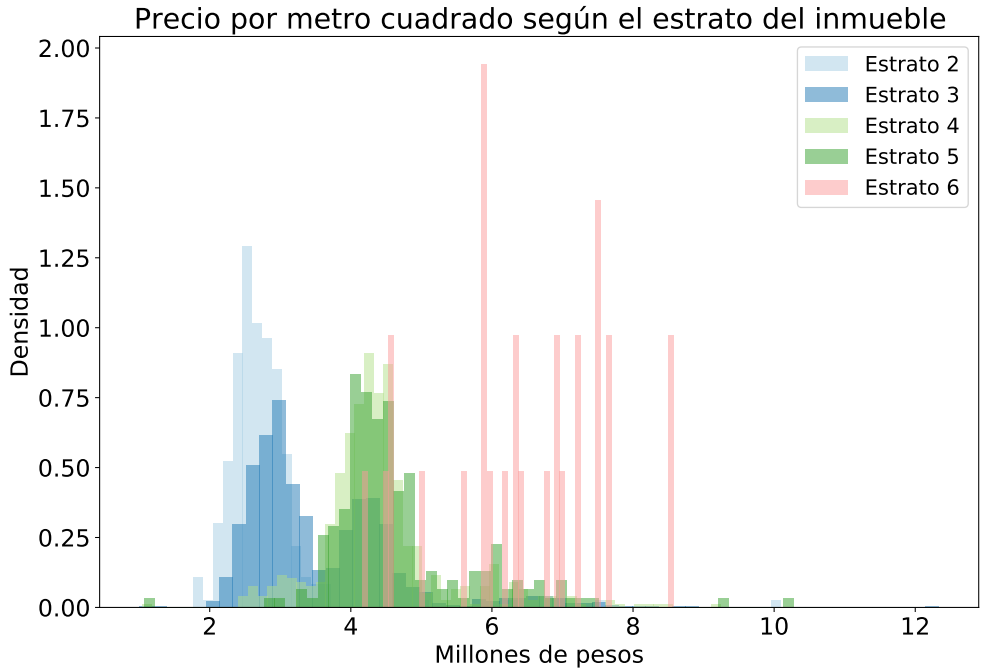


Para ahondar más en el análisis de la relación precio-área se calculó el precio solicitado por metro cuadrado de cada inmueble. La Tabla 3 muestra el promedio y la desviación estándar del valor por metro cuadrado de los inmuebles según su estrato y la Figura 4 muestra la distribución completa también por estrato. Como es de esperarse, también hay un aumento del precio por metro cuadrado de los inmuebles a medida que aumenta el estrato pero se observa una menor variabilidad del precio dentro de cada estrato. Hay una distribución bimodal de los precios, el promedio es cercano entre los inmuebles de estratos 2 y 3 y también las distribuciones de los estratos 4 y 5 están agrupadas. Parece que mirar el precio por metro cuadrado en vez del precio total da una variable menos volátil.

Tabla 3: Precios promedio según el estrato de los inmuebles

Estrato	Precio m^2 promedio (millones)	D.E.
2	2.755	0.703
3	3.596	1.127
4	4.430	0.916
5	4.680	1.041
6	6.437	1.181

Figura 4: Precio solicitado por metro cuadrado de los inmuebles según el estrato



2. **Estimación del valor de los inmuebles.** Es razonable asumir que cuando se pone en venta un inmueble, su precio estará sesgado hacia arriba, dado que los propietarios quieren obtener la mayor cantidad de ganancia posible. De esa manera, es importante poder extraer el valor real de la propiedad a la hora de evaluar si es rentable comprar un inmueble que se está ofertando. Para revelar ese precio "real" se puede hacer uso de las variables observables de los inmuebles para inferir un precio promedio que se ajuste a las características de cada inmueble y así decidir si el precio solicitado de venta es más alto o más bajo de lo que debería ser dadas las características de la propiedad.

   Para plantear el problema que se quiere solucionar, se puede plantear un modelo en el que el precio de un inmueble está formado a partir de las características físicas que tenga, de su ubicación y de un componente aleatorio distinto para cada inmueble. En términos generales, la relación se puede escribir de la siguiente manera

$$Precio_i = f\left(X, \varepsilon_i\right) \tag{1}$$

   Aquí el reto es estimar esa función $f(\cdot)$. En 1, $Precio_i$ corresponde al precio por metro cuadrado del inmueble i; X contiene una serie de variables de interés que se espera que tengan efecto en el proceso de formación de precios; y ε_i es el componente aleatorio no explicado que absorbe los demás factores que afectan el precio de un inmueble.

   En este caso, se asume que la función que genera los precios es de una relación lineal entre las variables observables que están contenidas en la muestra proporcionada para el ejercicio. En particular, la ecuación se escribiría de la siguiente manera

$$Precio_i = \beta_1 + \beta_2 Piso_i + \beta_3 Garaje_i + \beta_3 Antigüedad_i + \beta_4 Ascensor_i + \gamma UPZ_i + \delta Estrato_i + \varepsilon_i \tag{2}$$

   donde β_j son los coeficientes a estimar e indican cómo afecta cada característica observable al precio de un inmueble. $Piso_i$ es una variable numérica cuyo valor es el piso en el que está ubicada el inmueble i; $Garaje_i$ es una variable binaria que toma el valor de uno si el inmueble

i tiene garaje y cero de lo contrario; $Antigüedad_i$ es numérica y corresponde a los años que tiene el inmueble i; $Ascensor_i$ toma el valor de uno si el inmueble i tiene ascensor y cero de lo contrario; UPZ_i es una colección de variables binarias que indican en cuál de las 69 UPZ que están registradas en la base de datos está ubicada la propiedad y γ sería una serie de coeficientes que capturan la variabilidad en el precio que ocurre debido a características intrínsecas de las UPZ de la ciudad; y, finalmente, $Estrato_i$ contiene variables binarias que indican en qué estrato está ubicado cada inmueble.

La idea del modelo propuesto en 2 es estimar los coeficientes β y γ para poder encontrar la relación entre las características de un inmueble y su precio de venta y luego estimar con base en esas relaciones cuál sería el precio promedio de un inmueble con valores particulares de esas variables. Es decir, se va a estimar un precio "estructural" que estaría limpio de factores externos como el *mark up* de quién lo está vendiendo y que utiliza el precio de inmuebles de características similares para extraer el valor real de la propiedad.

La Tabla 4 muestra los coeficientes estimados de la ecuación 2 para la muestra que quedó tras el proceso de limpieza de datos. La primera columna tiene el nombre de cada variable explicativa, la segunda contiene el valor estimado del coeficiente para cada variable, la tercera contiene el p-valor del coeficiente estimado -con un valor menor a 0.05 se considera que el coeficiente estimado es significativo al 95 % de confianza- y las últimas dos columnas muestran los intervalos de confianza del estimador al 2.5 %. Además de las variables nombradas, se incluyeron efectos fijos de la UPZ y del estrato al que pertenece cada inmueble pero los 75 coeficientes no se muestran para ahorrar espacio, además de que no tienen una interpretación particularmente útil individualmente.

La regresión permite concluir que que el piso en el que está un inmueble no tiene un efecto significativo sobre su precio; un inmueble que tenga garaje cuesta en promedio 457 mil pesos mayor; cada año de antigüedad reduce a un inmueble 30 mil pesos; un inmueble que tenga ascensor tiene un precio en promedio 204 mil pesos mayor; y el precio promedio por metro cuadrado para los inmuebles de la muestra, tras controlar por estas variables, es de 3'098 mil pesos.

Tabla 4: Regresión lineal del precio por metro cuadrado contra las características de los inmuebles

	Coeficiente	P-valor	[0.025	0.975]
Intercepto	3.098	0.0	2.247	3.95
	(0.245)			
Piso	0.0	0.939	-0.01	0.012
	(0.006)			
Garaje	0.457	0.0	0.366	0.547
	(0.054)			
Antiguedad	-0.03	0.0	-0.03	-0.02
	(0.003)			
Ascensores	0.204	0.0	0.106	0.302
	(0.048)			

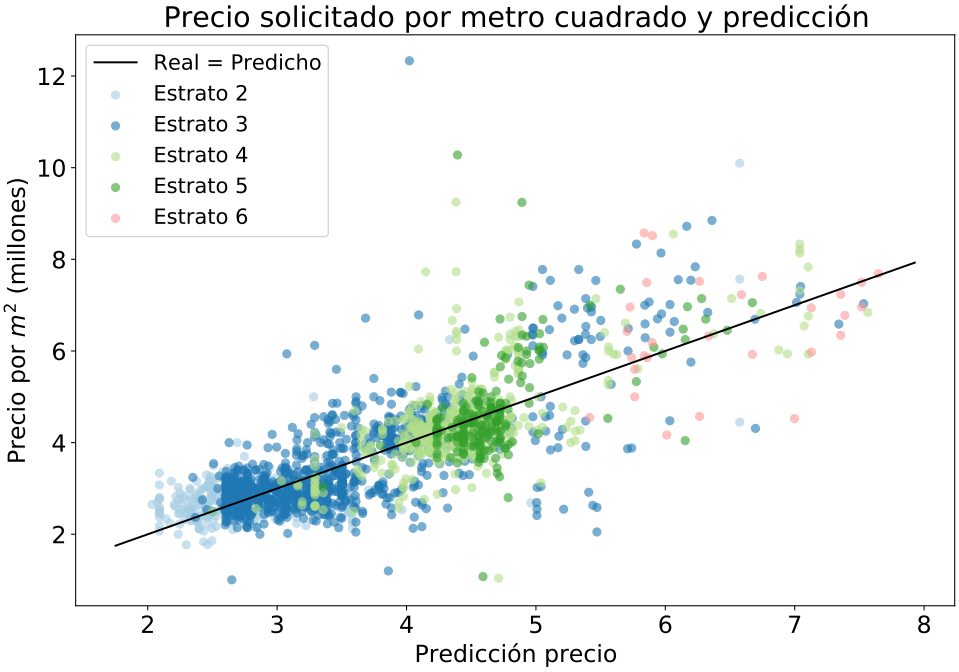
Nota: Errores estándar entre paréntesis. Se incluyeron efectos fijos por UPZ y por estrato.

Tras estimar estos coeficientes, podemos meter los valores de cada variable en 2 para calcular una predicción de $Precio_i$ y comparar este precio, que se debe acercar al valor real de los inmuebles, con el precio de venta para analizar cuáles inmuebles están sobrevalorados. La Figura 5 muestra esa relación. En el eje vertical está el precio original por metro cuadrado solicitado para cada inmueble y en el horizontal está la predicción calculada a partir de la regresión lineal descrita anteriormente. Cada punto tiene un color que corresponde al estrato al que pertenece el inmueble y la línea negra es una de 45 grados que permite separar las

propiedades que están “sobrevaloradas” y las que están “subvaloradas”.

Aquellos inmuebles que tienen un precio de venta demasiado alto están por encima de la línea negra, pues su precio solicitado es mayor al valor real estimado; y, por el otro lado, si están por debajo de la línea negra es porque su precio de venta es menor al valor real estimado de la propiedad. Es notable que para algunos de los precios por metro cuadrado más altos -los puntos de más arriba en el eje vertical- tienen valores predichos cercanos a la mitad de la distribución horizontalmente. Esto quiere decir que son demasiado caros para un inmueble con sus características físicas y su ubicación, por lo que la predicción de la regresión lineal los valora mucho menos que su precio de venta.

Figura 5: Diagrama de dispersión del precio por metro cuadrado solicitado y el precio predicho



Lo que vimos hasta ahora fue una manera relativamente sencilla de extraer el valor real de un inmueble a partir de sus características físicas y del precio de venta que se solicita. De manera que se puede saber cuándo un inmueble está siendo vendido por un precio mayor a su valor real -de modo que comprarlo no es rentable- y cuándo está siendo vendido por un valor menor a su valor real.

La metodología de regresión lineal es relativamente sencilla pero también es muy simple y fácil de interpretar. Si se quisiera profundizar más en se pueden aplicar herramientas de *Machine Learning*, como redes neuronales o *Gradient Boosting*, que permiten estimar la función $f(\cdot)$ con otras formas más allá de la lineal y encontrarían potenciales relaciones mucho más complicadas entre las variables del modelo. También se podría incluir información georeferenciada de la ubicación de los inmuebles, pues hay muchos factores, como acceso al servicio de transporte público, seguridad o presencia de parques, que pueden afectar mucho el precio de un inmueble más allá de las características físicas incluídas en el análisis hecho anteriormente.