

Ejercicios para practicar I

El conjunto de datos VentaViviendas contiene información sobre el precio de venta de una serie de viviendas, junto con las características básicas de las mismas. Las variables contenidas en el fichero son (observa que hay dos variables objetivo diferentes):

Variable	Descripción
Year, month	Año y mes de la venta
Price (objetivo)	Precio de venta de la vivienda
Luxury (objetivo)	Variable dicotómica que toma valor 1 si se trata de una vivienda de lujo (precio superior a medio millón de \$) y 0, en caso contrario
bedrooms	Número de habitaciones
bathrooms	Número de baños (los medios se refieren a aseos)
sqft_living	Superficie del salón
sqft_lot	Superficie total (incluye el jardín)
sqft_above	Superficie excluyendo el sótano
basement	¿Tiene sótano? (1: sí, 0: no)
floors	Número de plantas
waterfront	¿Tiene vistas al mar? (1: sí, 0: no)
view	¿Tiene buenas vistas? (1: sí, 0: no)
condition	Estado de la vivienda (de A a D, siendo A el mejor estado)
yr_built	Año de construcción de la vivienda
yr_renovated	Año de renovación de la vivienda (si es 0, no ha sido renovada)
lat, long	Coordenadas de latitud y longitud de la vivienda

El ejercicio consta de los siguientes apartados:

- 1) Importa el conjunto de datos y verifica que el tipo de las variables se haya asignado correctamente.
- 2) Realiza un análisis descriptivo para determinar si existen errores en las variables (valores mal codificados, valores fuera de rango, categorías con poca representación, simetría y curtosis de las variables cuantitativas, etc.).
- 3) Corrige los errores detectados en el apartado anterior.
- 4) Crea una variable categórica a partir de la variable "yr_renovated" que tome el valor 0 cuando la variable no haya sido renovada; y 1, en otro caso. Rechaza la variable original.
- 5) Realiza un estudio sobre los datos atípicos y transforma los que encuentres en valores faltantes. Recuerda verificar que la cantidad de atípicos no sea excesiva.
- 6) Trata los datos ausentes, lo que implica verificar si hay alguna variable y/o observación que tenga demasiados y si existe algún patrón que justifique su presencia, imputar las variables que los tengan y crear una variable que recoja la información de las imputaciones.