Ejercicios para practicar: Depuración

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:

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 salon
sqft_lot	Superficie total (incluye jardín)
Sqft_above	Superficie excluyendo el sótano
floors	Número de plantas (los medios se refieren a entre plantas)
waterfron	¿Tiene vistas al mar? (1:si, 0: no)
view	¿Tiene buenas vistas? (1:si, 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)
basement	¿Tiene sótano? (1: sí, 0: no)
lat, long	Coordenadas de latitud y longitud de la vivienda

El ejercicio consta de las siguientes partes:

1) Crear un proyecto, instala las librerías necesarias y cárgalas.
2) Importar el conjunto de datos y asegúrate de que todas las variables sean del tipo adecuado.
3) 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.).
4) Realiza las correcciones pertinentes, teniendo en cuenta los errores detectados con anterioridad y verifica que los cambios se hayan realizado correctamente.
5) 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.
6) Analiza la existencia de valores atípicos para las variables cuantitativas según el método más apropiado y transformarlos a datos faltantes si lo consideras oportuno.
7) Crea una variable que guarde información sobre la proporción de valores perdidos por observación.
8) Analiza si existe alguna observación y/o variable con demasiados datos faltantes y, de ser así, elimínala.
9) Para las variables con datos faltantes, decide cómo tratarlos, justificando la respuesta. En caso de imputar los datos ausentes. Hazlo con la media para las variables cuantitativas y de modo aleatorio con las variables cualitativas.