Ejercicios para practicar II

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 sotano? (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	Coordinadas de latitud y longitud de la vivienda

Partiendo del conjunto de datos que depuraste en la última clase, el objetivo final de estos ejercicios es construir un modelo de regresión lineal para predecir la variable *Price*. Los ejercicios constan de los siguientes apartados:

1) Crea una variable aleatoria.
2) Determina cuáles serán las variables más útiles para predecir cada una de las variables objetivo a partir de los gráficos y el valor de la V de Cramer.
3) Transformas las variables continuas de input de manera que se maximice la relación con las variables objetivo por separado. ¿Se aplican las mismas transformaciones para las dos variables objetivo?
4) Realiza una partición *Entrenamiento-Prueba* (80-20) de los datos.
5) Construye un primer modelo de regresión lineal en el que incluyas todas las variables disponibles (sin las transformaciones automáticas ni las interacciones pero incluyendo las variables aleatorias). Evalúa la calidad del modelo resultante e interpreta el parámetro de una variable continua y otra binaria.
6) Basándote en los resultados del apartado 2, construye un modelo de regresión que contenga únicamente las variables detectadas. ¿Este modelo es mejor que el anterior?
7) Basándote en la importancia de las variables del modelo *inicial*, determina las variables menos útiles para predecir el precio de la vivienda. A continuación, construye un modelo de regresión como el del apartado 5 pero eliminando las variables detectadas. ¿Este modelo es mejor que los anteriores?
8) Partiendo del mejor modelo de los 3 anteriores (ten en cuenta su comportamiento en entrenamiento y prueba, así como el número de parámetros que tienen), incluye las interacciones que consideres puedan ser influyentes y determina si lo son o no.
9) Utilizando validación cruzada (20 repeticiones, 5 grupos), determina cuál de los 4 modelos anteriores es preferible.
10) Evalúa el modelo ganador (estabilidad y bondad del mismo, variables más importantes, etc.).