Ejercicios para practicar IV

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 depurado, los ejercicios constan de los siguientes apartados:

1) Realiza una partición *Entrenamiento-Prueba* (80-20) de los datos.
2) Construye de nuevo el modelo ganador del segundo día de clase. Analiza los resultados y determina si existe alguna variable cualitativa cuyas categorías deban unirse. De ser así, crea una variable nueva con menos categorías y genera de nuevo el modelo pero cambiando dicha variable. ¿Observas alguna mejoría en el modelo?
3) Determina el mejor modelo de regresión lineal a partir de stepwise y backward basándote en los estadísticos AIC y el SBC incluyendo todas las variables disponibles (sin las transformaciones automáticas ni las interacciones). ¿Coinciden algunos de los 4 modelos generados? De no ser así, ¿cuál parece ser el mejor de todos?
4) Repite el ejercicio 3 (únicamente con stepwise), pero incluyendo todas las interacciones posibles.
5) Repite el ejercicio 3 (únicamente con stepwise), pero considerando las variables originales junto con las transformaciones automáticas.
6) Repite de nuevo el ejercicio 3 (únicamente con stepwise), incluyendo todas las variables, sus transformaciones automáticas y las interacciones.
7) Llegados a este punto, habrás obtenido varios modelos diferentes, es el momento de decidir cuál de todos ellos es preferible. Aplica validación cruzada repetida sobre todos ellos y determina cuál es el mejor de todos basándote en los resultados de los diagramas de cajas.
8) Lleva a cabo una selección de variables aleatoria con todas las variables (incluidas las transformaciones y las interacciones) y determina los dos modelos que más se repitan. ¿Son mejores que los modelos previamente obtenidos en el conjunto de prueba?

9) Construye un modelo de regresión lineal LASSO y determina el mejor valor de su parámetro. ¿Es mejor que los modelos previamente obtenidos en el conjunto de prueba?
10) Lleva a cabo validación cruzada repetida para determinar el mejor modelo de entre el ganador del apartado 7 y los construidos en los apartados 8 y 9.
11) Si te da tiempo, repite los ejercicios 2-10 pero, en lugar de trabajar con regresión lineal para predecir la variable *Price*, construye modelos de regresión logística para la variable *Luxury*.