Class Activity 2

The Paris Housing Prices data set is an artificially contrived data set that supposedly contains housing prices in Paris for the year 2021 (the target), as well as artificial data relating to various different attributes of the houses (shown below). You may assume that all data values are correct, even though some may appear non-sensical:

* SQM - size of house in square metres
* Bedrooms - number of bedrooms in house
* HasYard - does house include yard?
* HasPool - does house include a pool?
* Floors - number of floors the house has
* SuburbType – type of suburb the house is situated in (working/middle/upper class, or exclusive)
* NumPrevOwners - number of previous owners
* Made – year built
* HasStormProtector – does house have storm protector? 0=no, 1=yes
* Basement – size of basement in square metres
* Attic – size of attic in square metres
* Garages – number of garages the house has
* HasStorageRoom – does house have a storage room? 0=no, 1=yes
* LivingRooms - number of living rooms in house

You are required to perform the following tasks, using this dataset:

1. Import the dataset in to R-studio (remember to copy the code and paste it in to your R-script so that it can be saved):
2. Check the class of the imported dataset (convert to a data frame if need be)
3. Use appropriate functions to print the number of rows (observations) and columns (attributes) in the raw data frame
4. Remove any observations with missing values. Take note of the revised sample size.
5. Check the class of ALL the variables in the dataset using a single function
6. Practise clearing the console
7. Convert all character variables to factor variables:
8. Explore the distribution of the target variable using appropriate functions
9. Use an appropriate function from the Stats package to fit an OLS regression model to the entire data frame (i.e. using a statistical approach rather than a machine learning approach)
10. Apply an appropriate function to observe the estimated regression coefficients, and model diagnostics:
11. Would you describe the bias of the fitted OLS model as being high or low?
12. Do the relationships in the data truly appear to be non-linear?
13. Interpret the parameter estimates for the qualitative variables.
14. Next, perform some additional data preprocessing…
15. Convert all factor variables to a numeric format using the “ifelse” function:
    1. Ensure that binary variables are coded 1 for yes, and 0 for no
    2. Convert SuburbType in to a numeric variable with ordered levels going from 1 to 4, where 4 corresponds with most exclusive
16. Create an additional attribute which describes the age of the house in years, in the year 2021
17. Re-check the class of attributes to ensure that they are all now in a numeric format. Make any final necessary conversions.
18. Create another data frame with a different name which excludes the “Made” attribute (use the “select” function from the dplyr package).
19. Fit an OLS model to the new data frame containing only numeric attributes
    1. Which parameter estimates are significant @5% level?
    2. Which parameter estimates are significant @20% level?
20. Fit a reduced OLS model using only those attributes whose parameter estimates were significant @20%. Compare the model diagnostics between the OLS models fitted in step 6 and step 7.