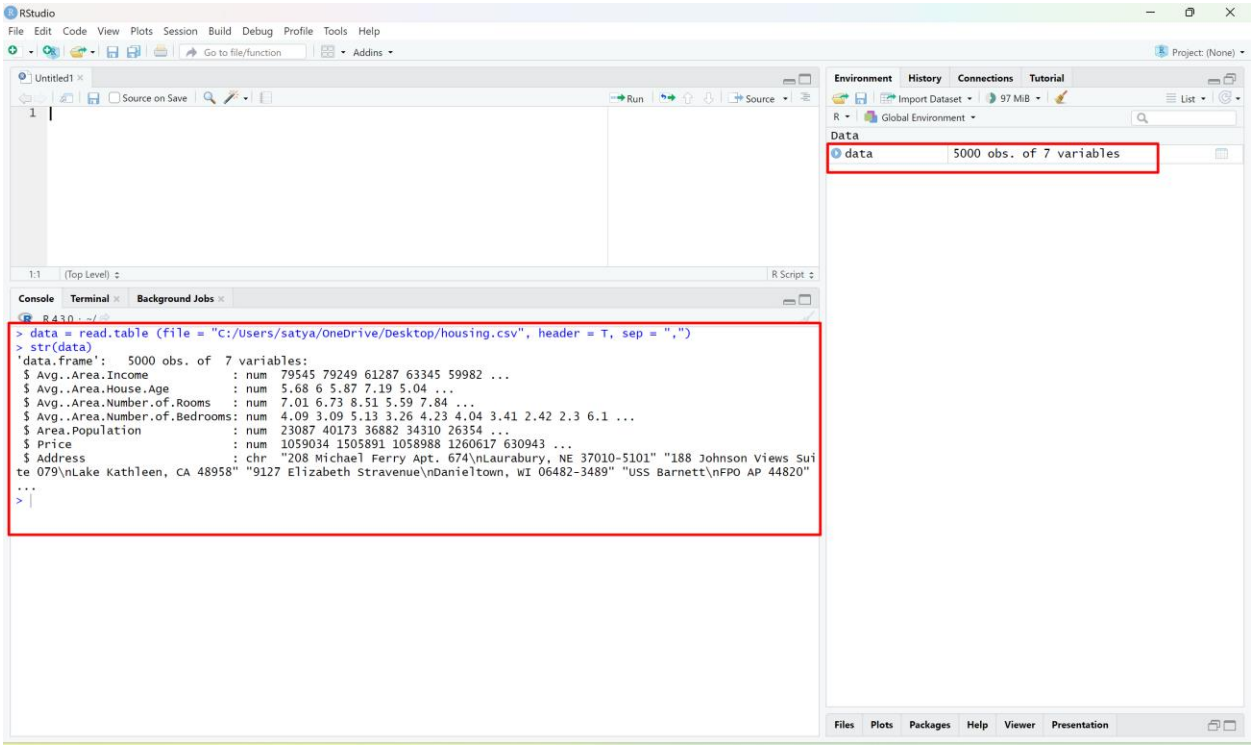
Key Outputs

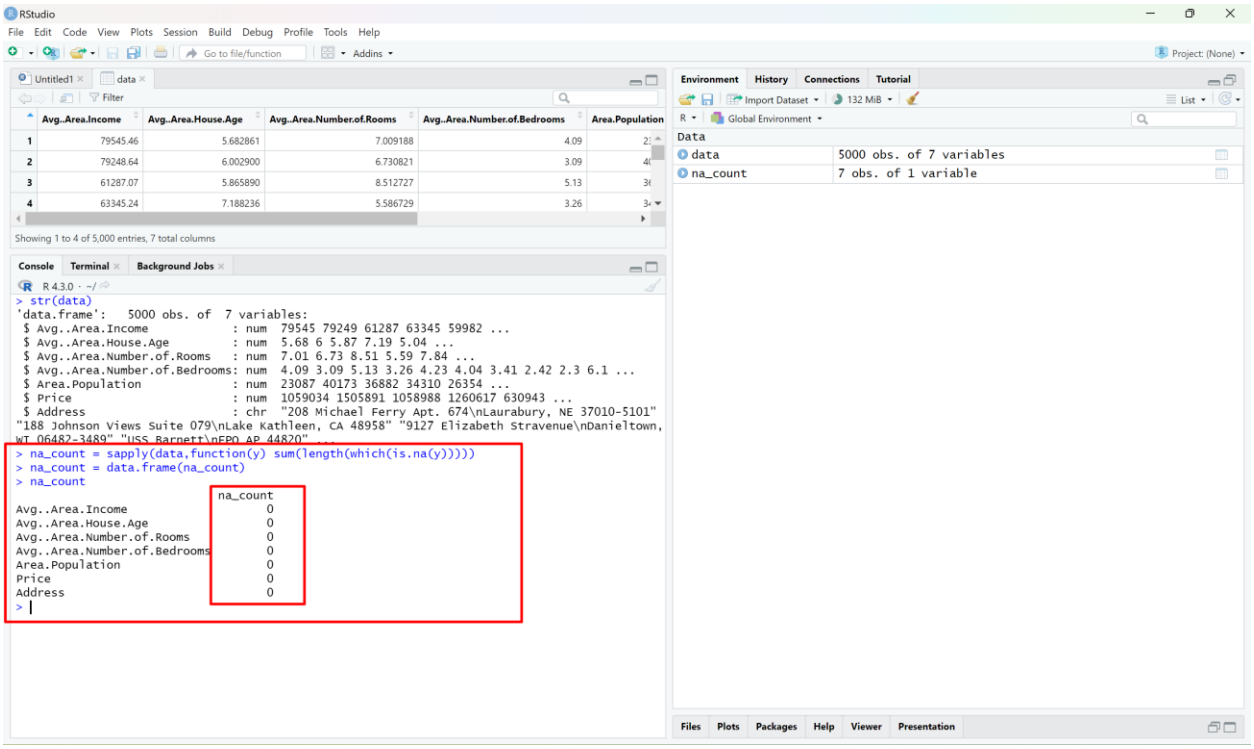
Name: Naga Satya Silpa Annadevara

A20517818

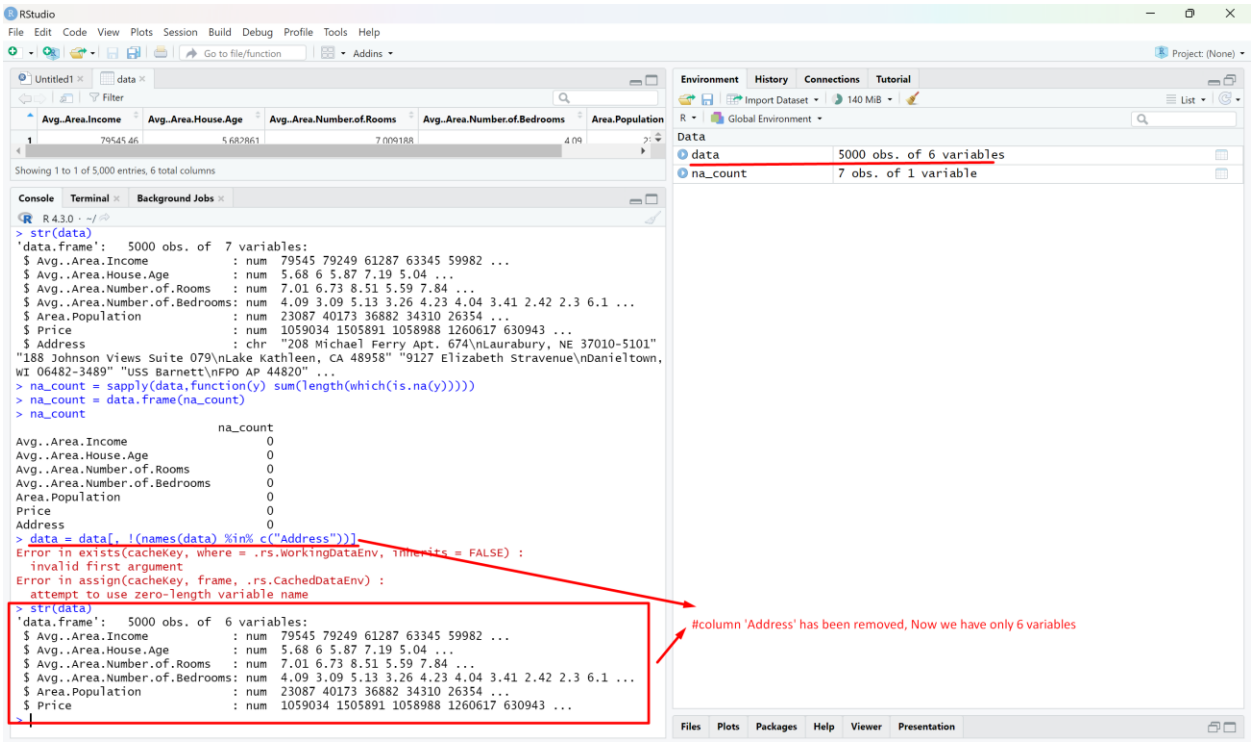
Load the housing.csv data into R:



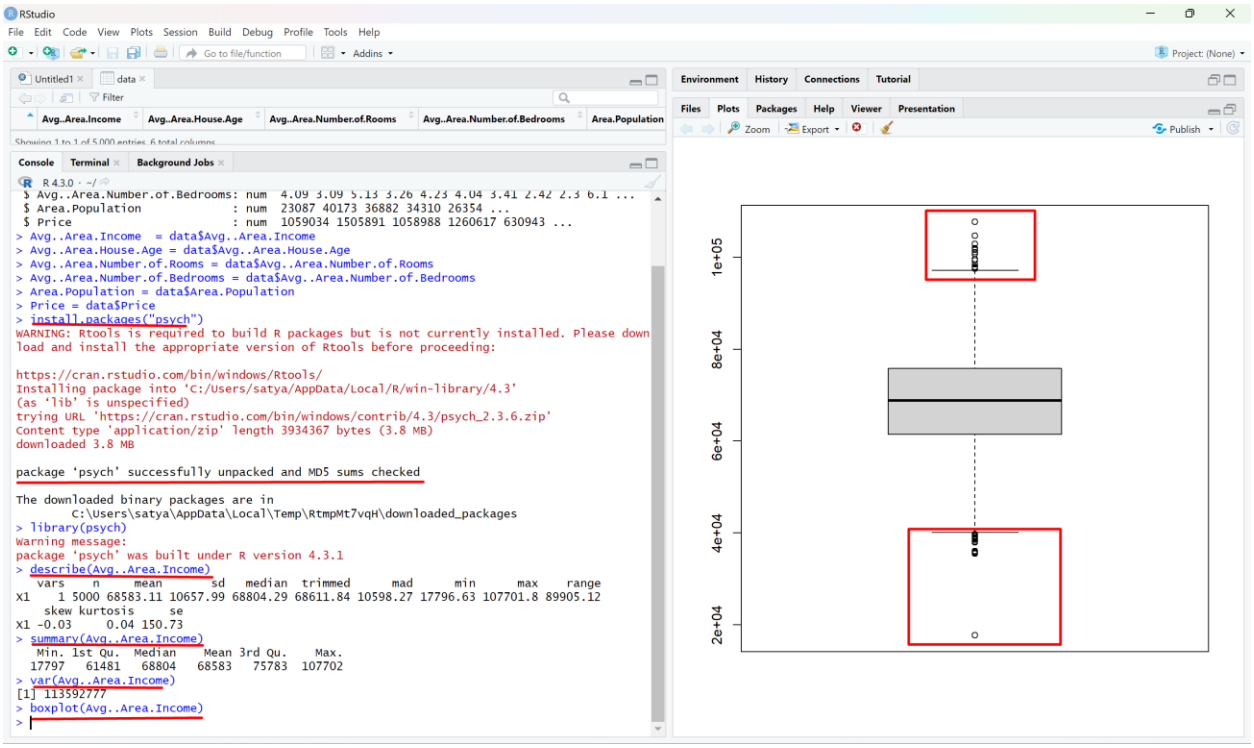
1.checked for missing values

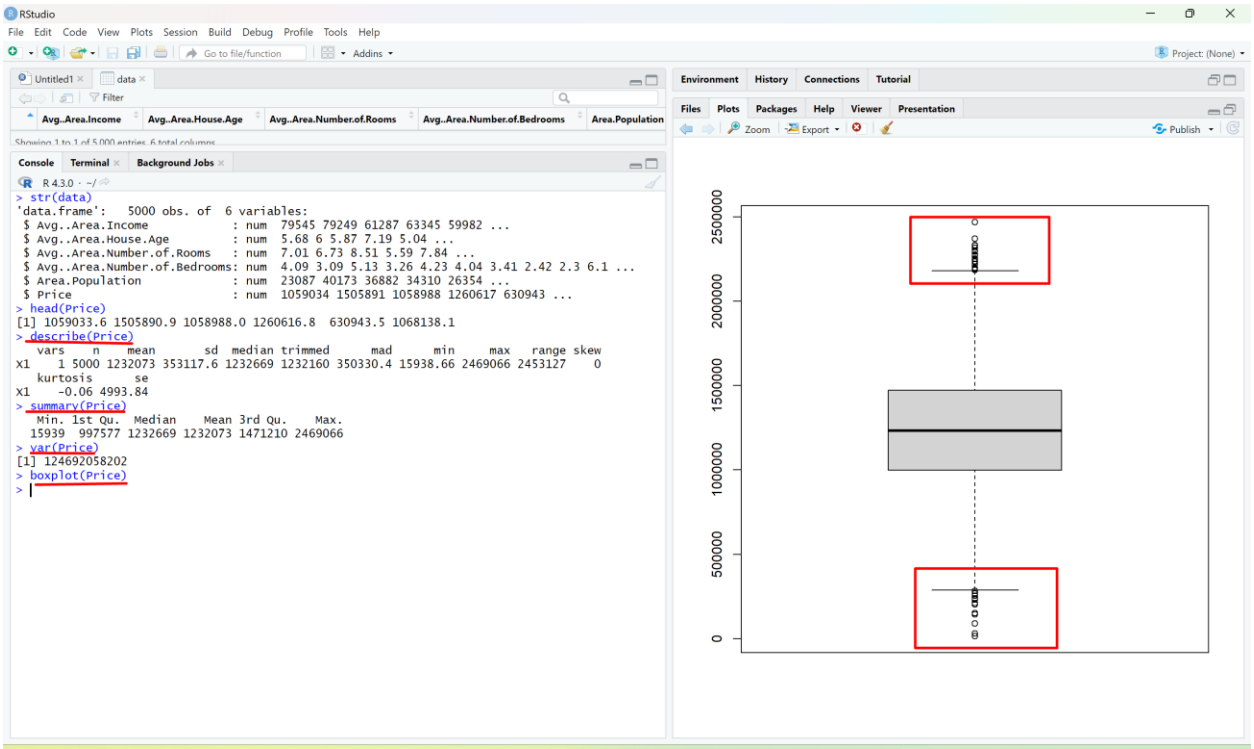


2. Then handled non-numeric data. Removed column 'Address':

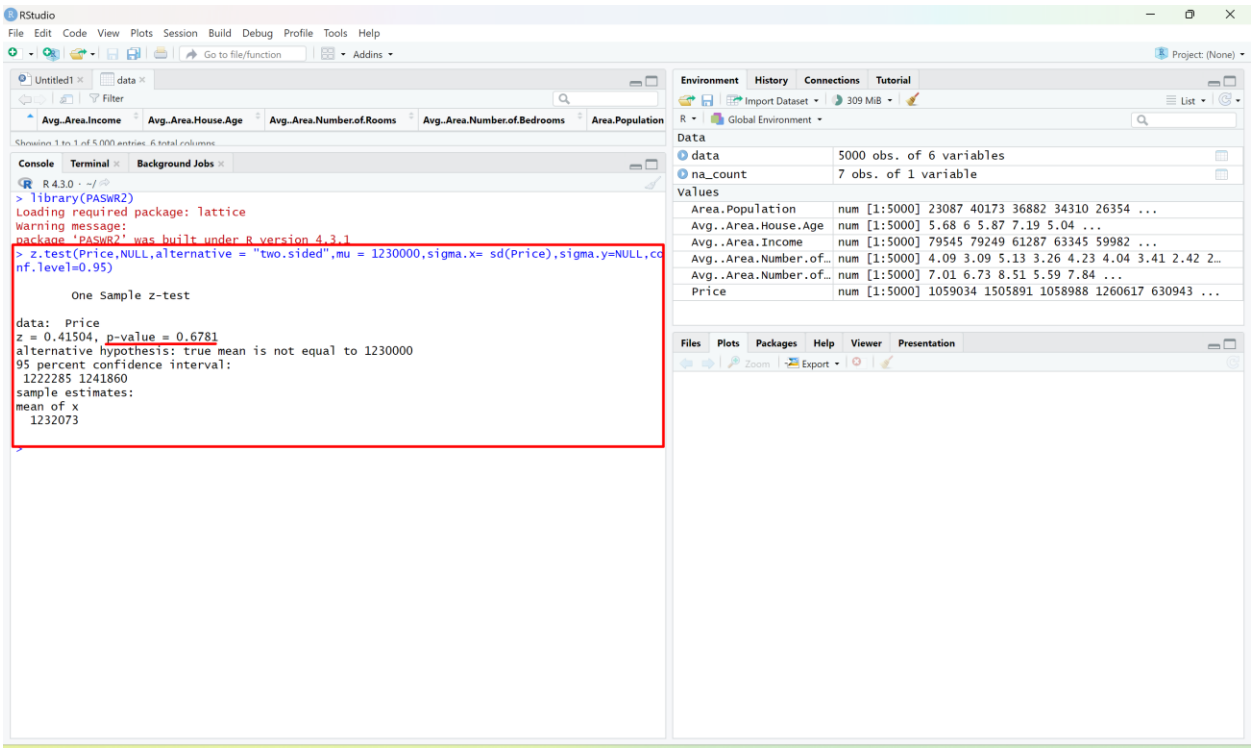


3. Performed Descriptive statistics on columns ‘Price & Income’:

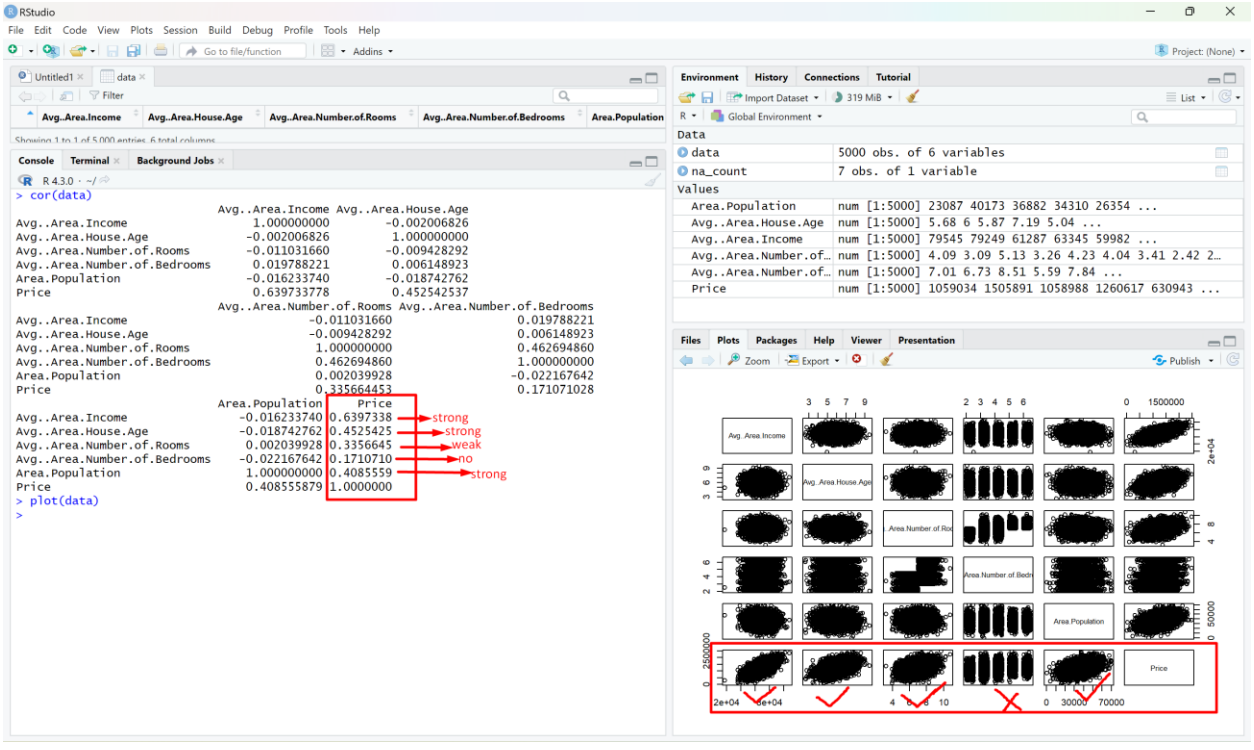




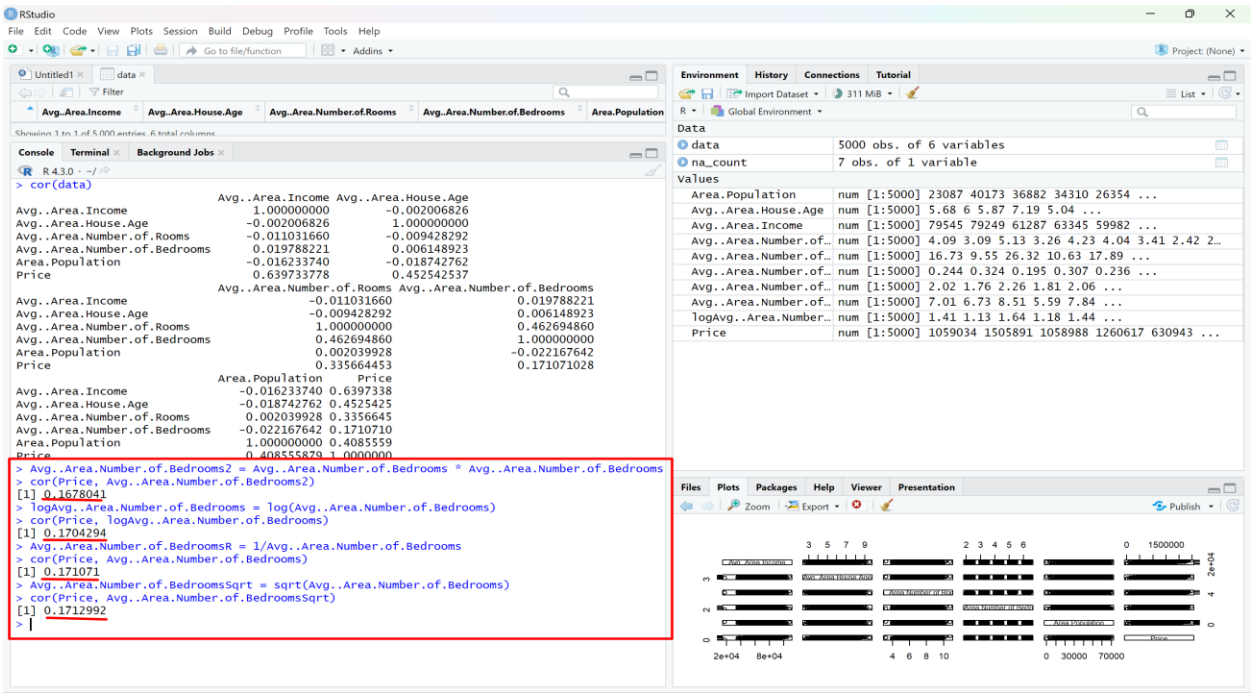
4. Validate the claim/Assumption by using **Hypothesis testing:**



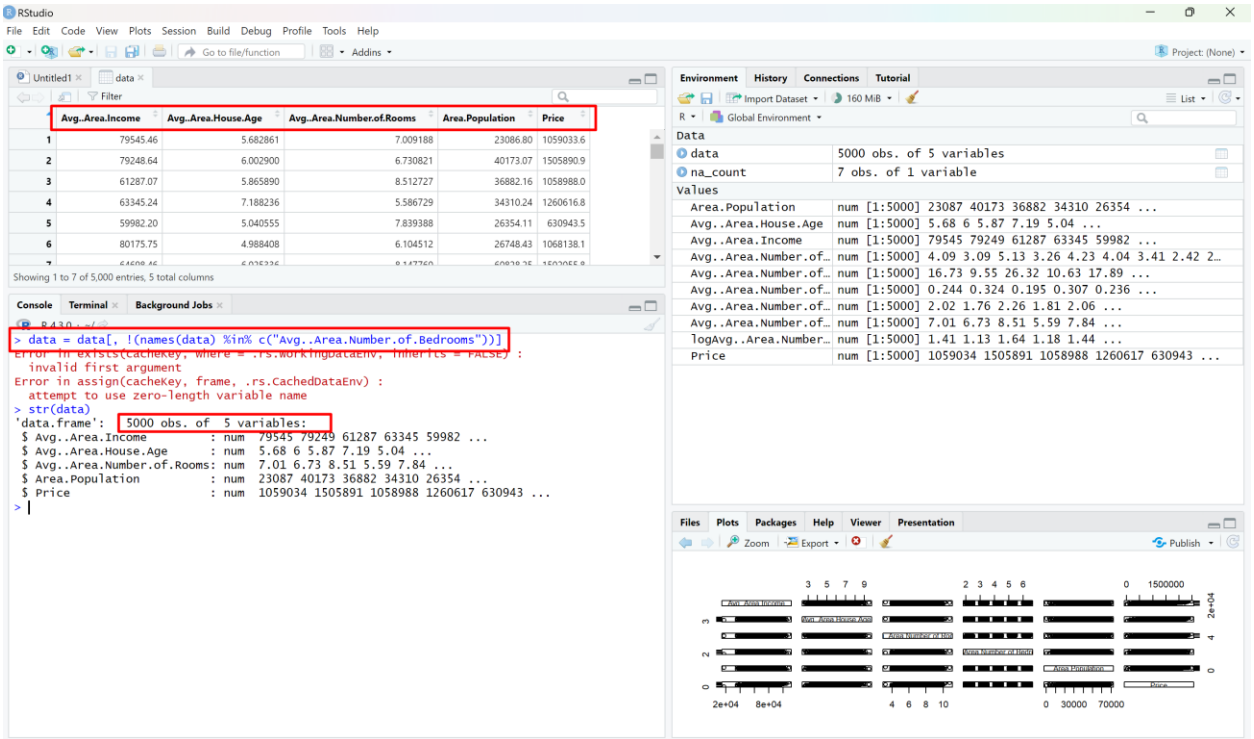
5. Handling non-linearity: Examine linear relation b/w X & y variables (check correlations)



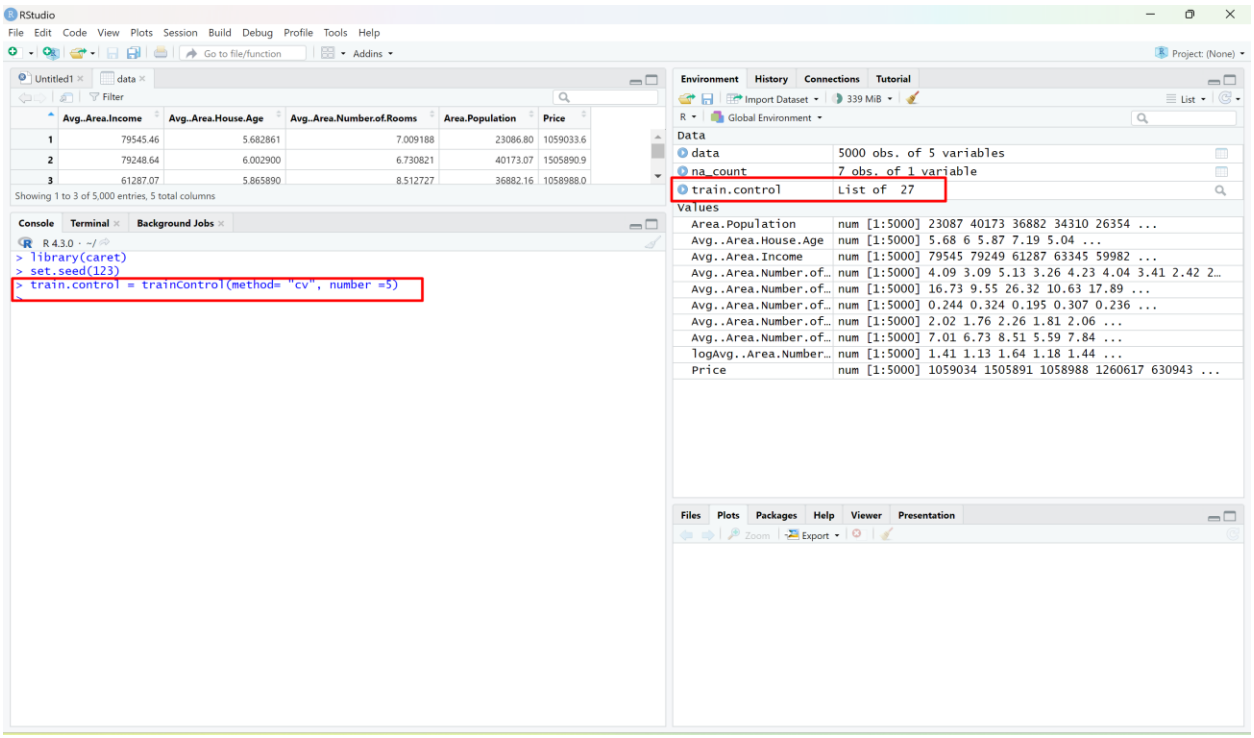
Performed transformation on X-variable: Avg..Area.Number.of.Bedrooms



droping the x-variable: Avg..Area.Number.of.Bedrooms:

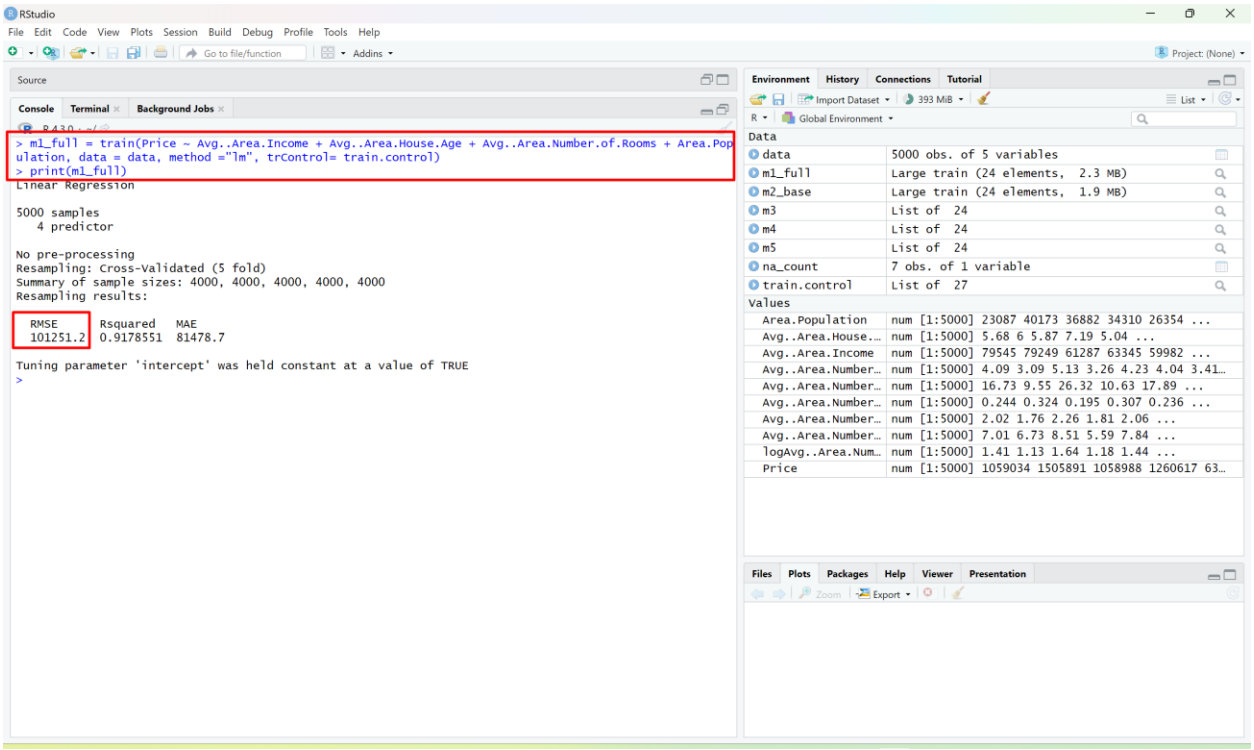


6. data split by N-fold cross evaluation:

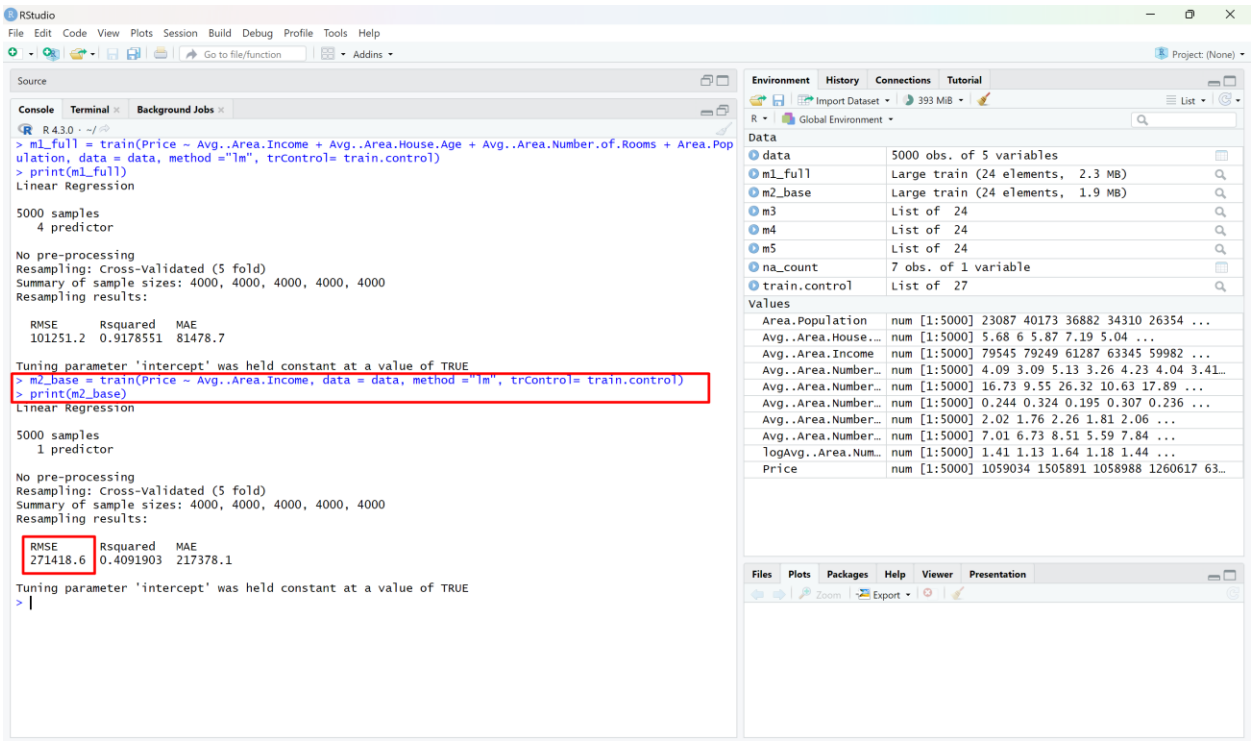


7.Build 5 different multiple linear regression models to compare them:

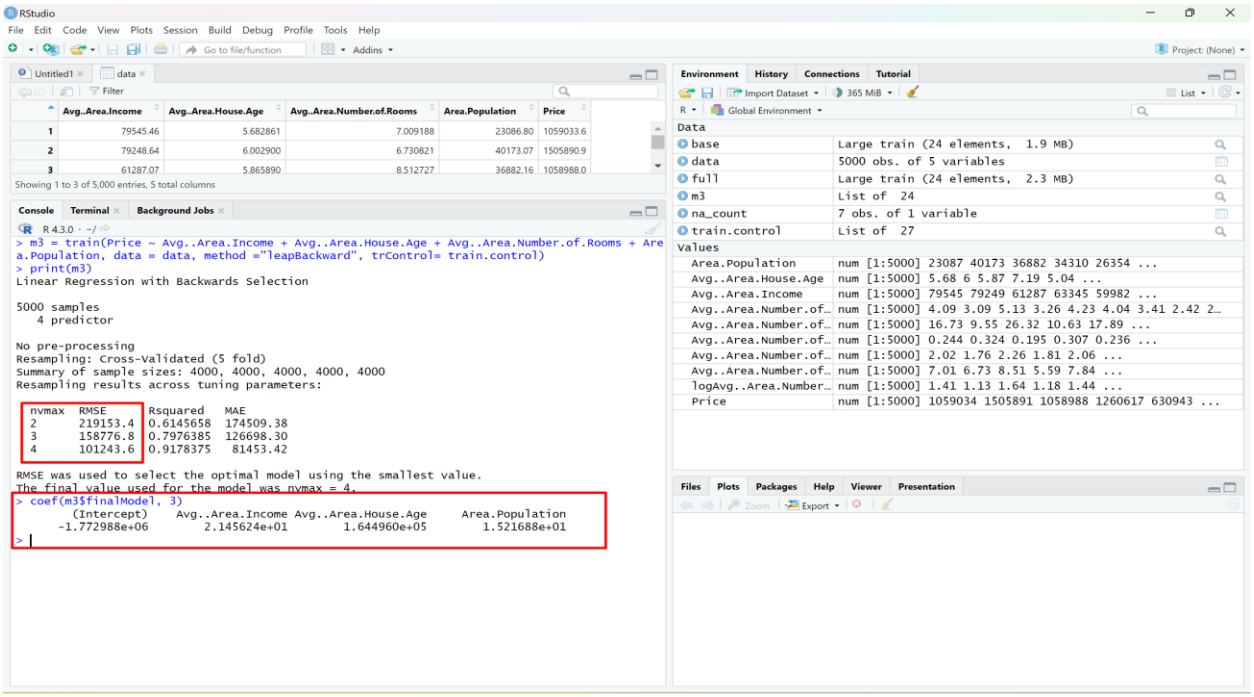
Building **Full model** by using all x-variables & target variable: **M1**



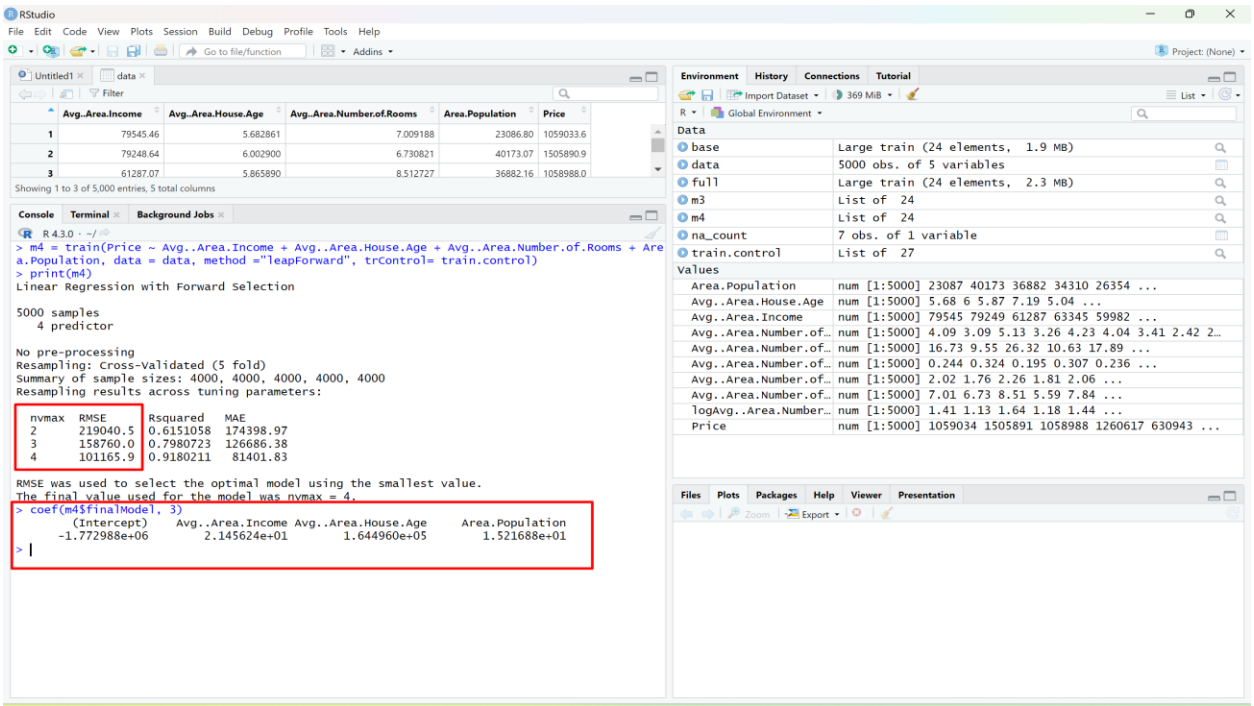
Building **base model** with just one x-variable & target variable:**M2**



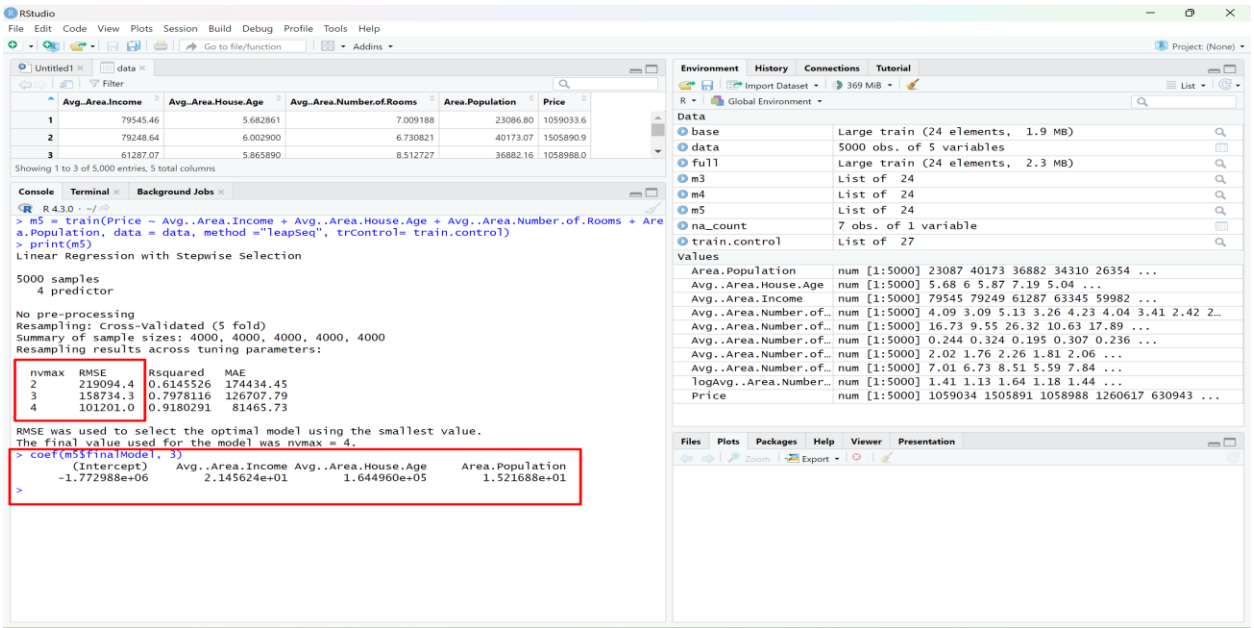
Building m3 by backward method: **M3**



Building m4 by forward method: **M4**

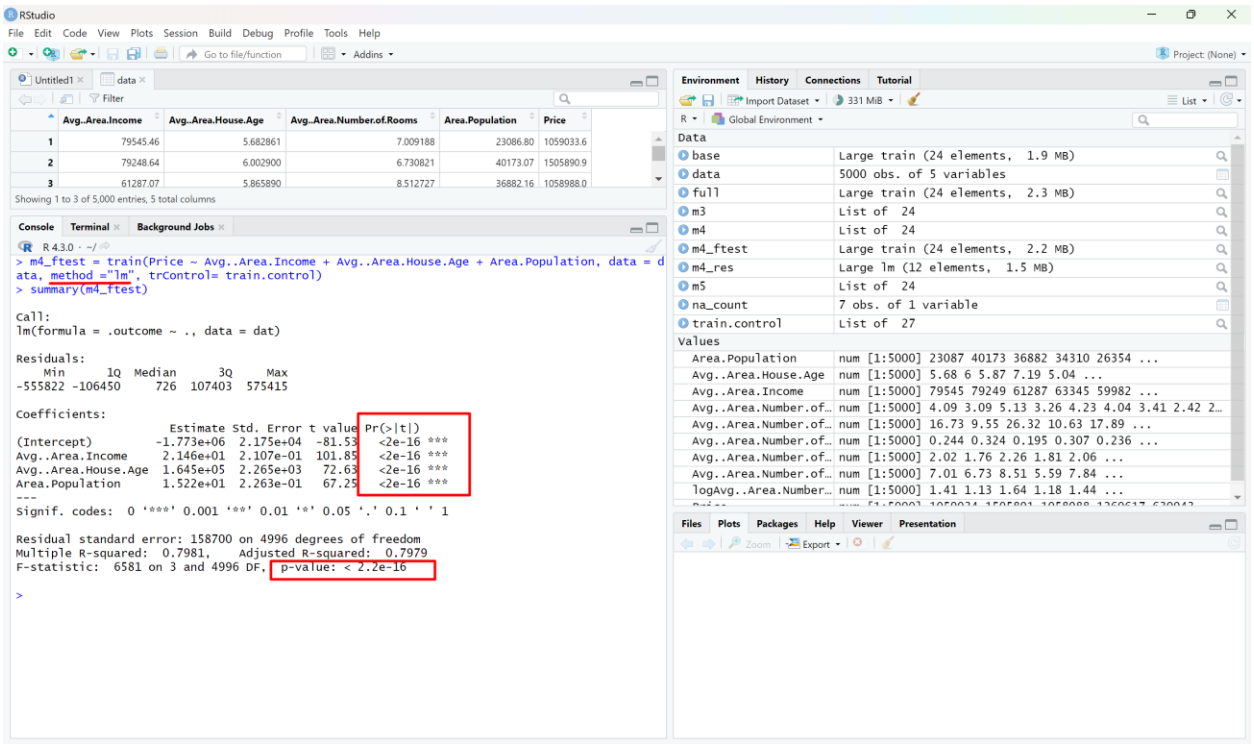


building m5 by forward method: **M5**

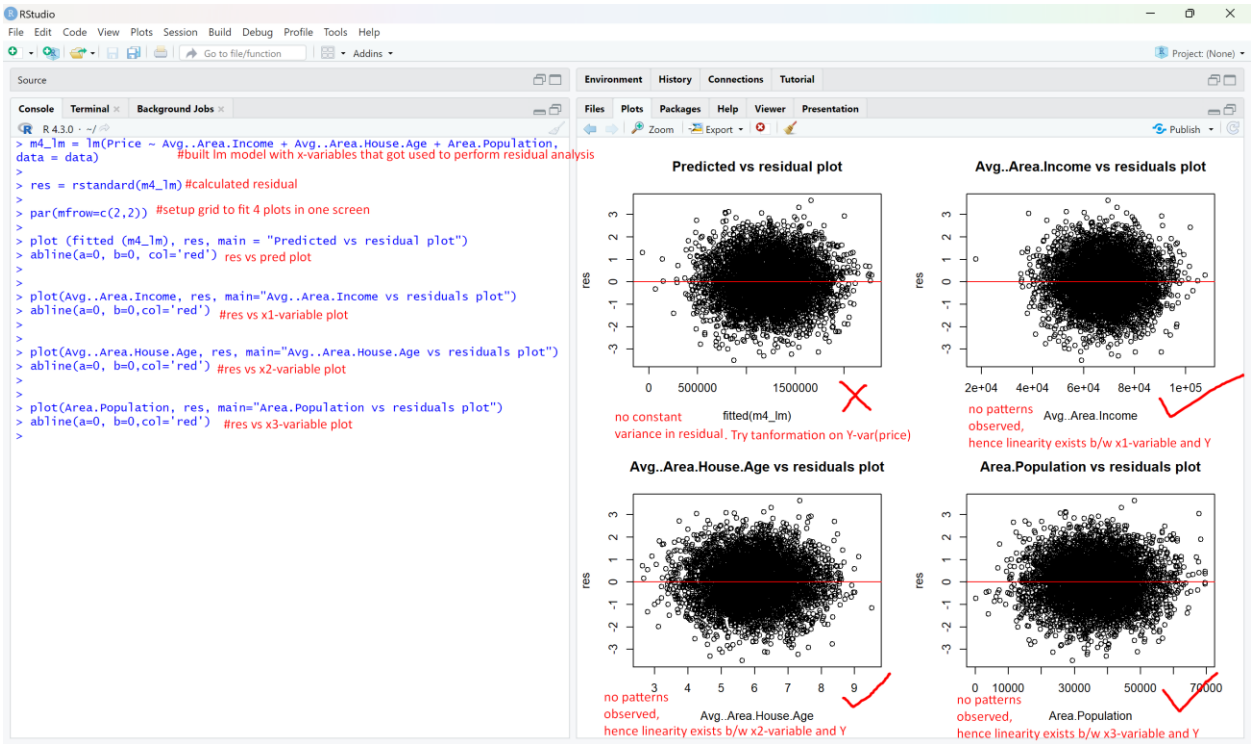


7. Model diagnosis:

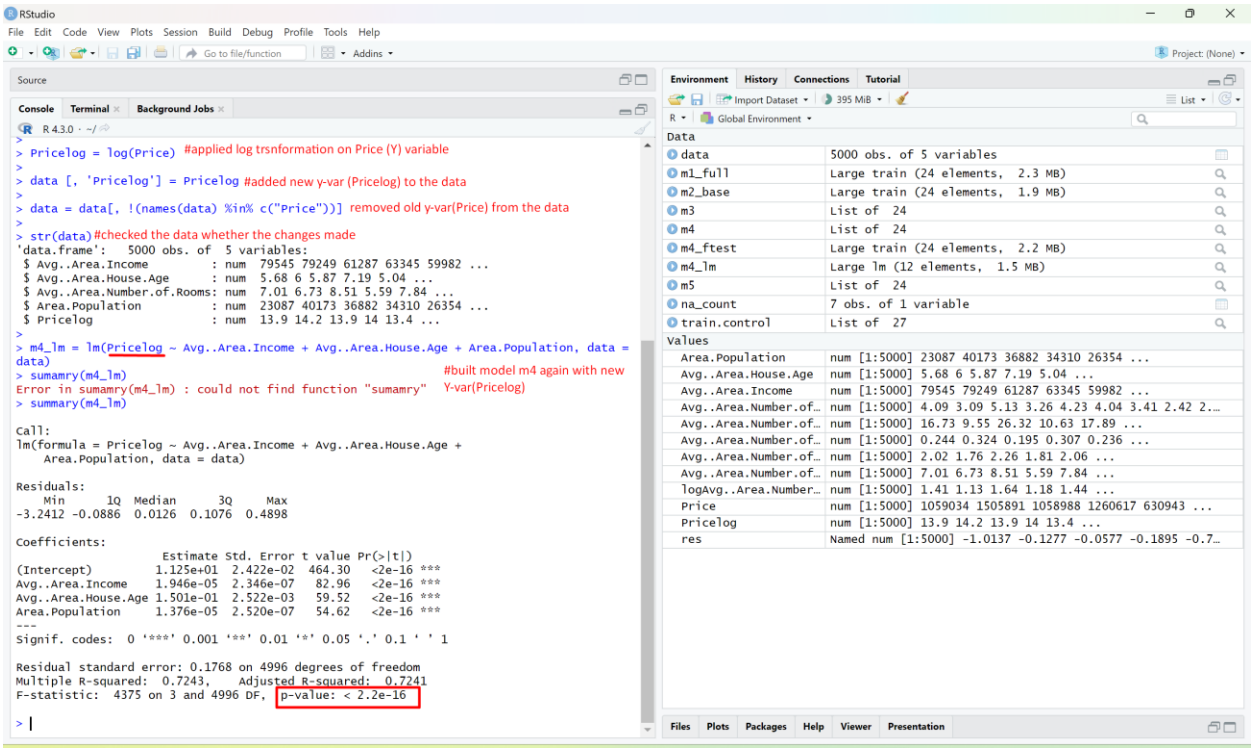
F-test on M4:



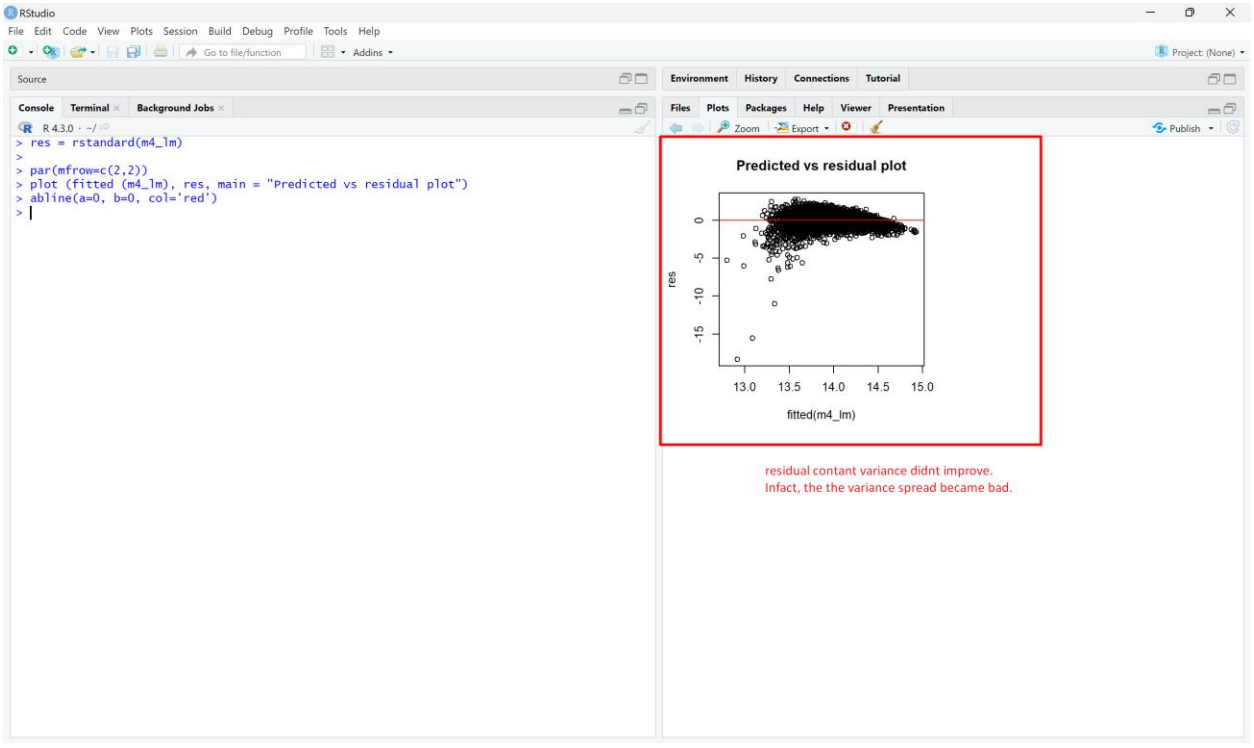
Residual analysis on M4:



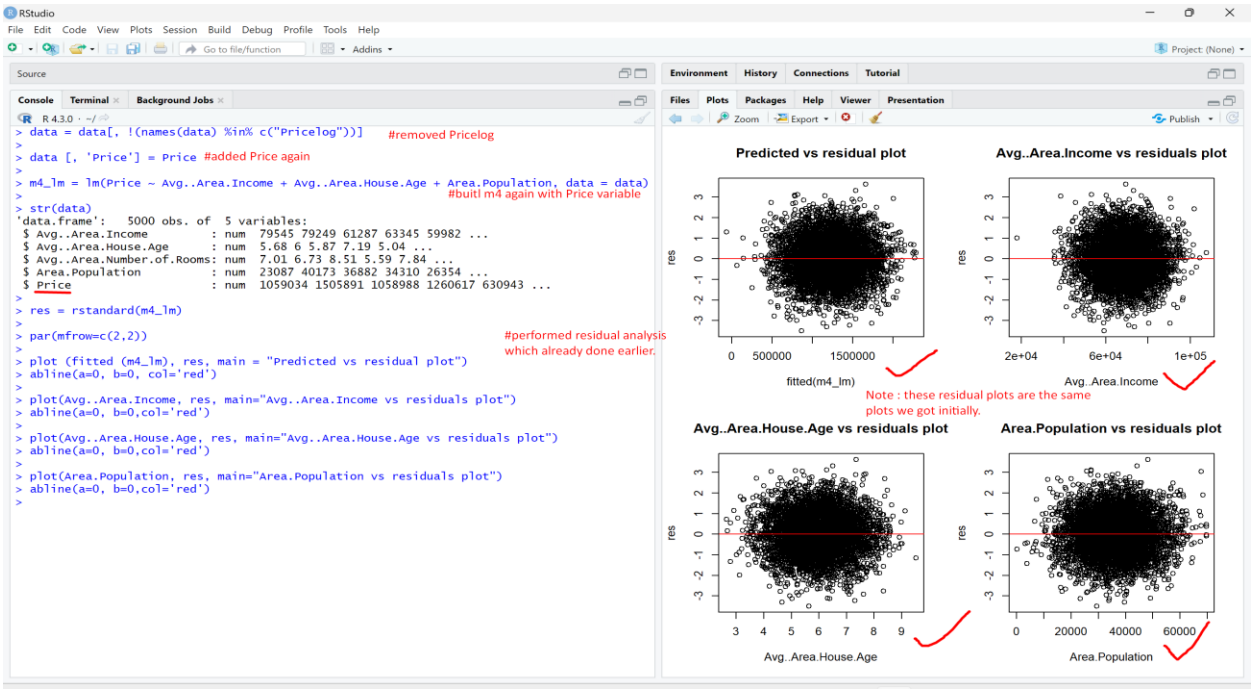
Tranformation on Y-variable & rebuilding the model again:



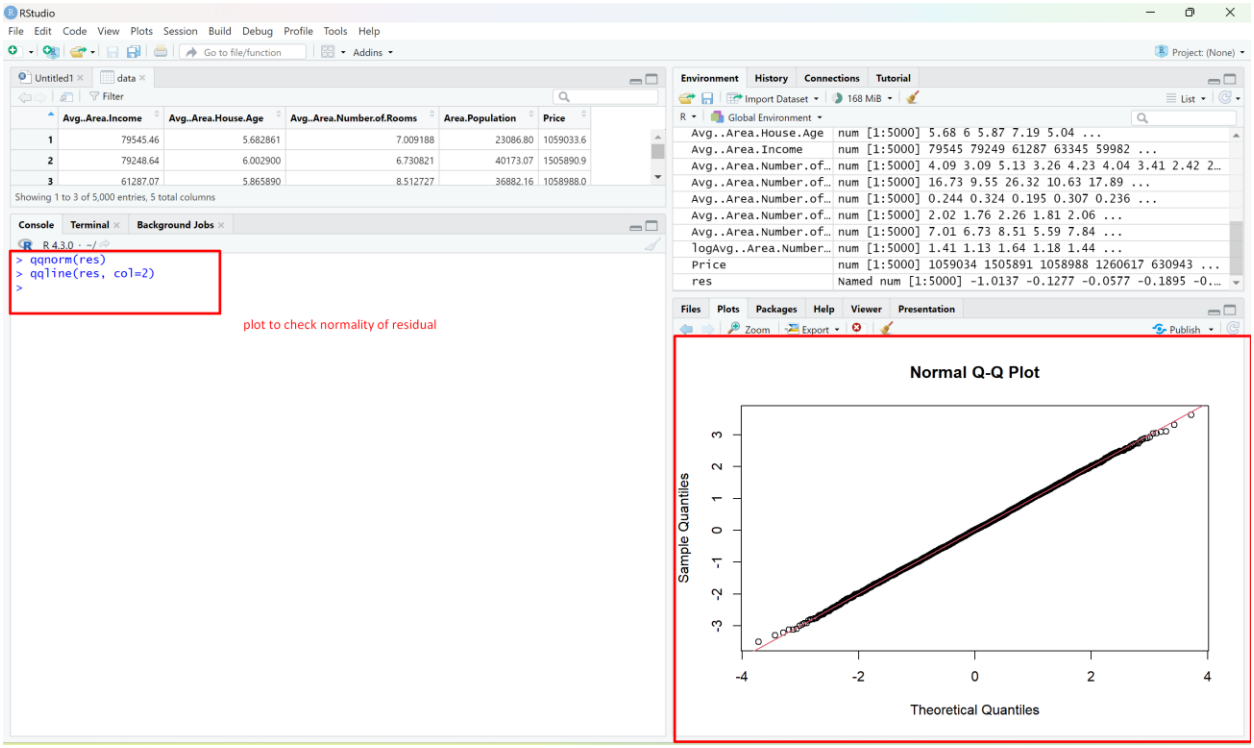
Re-perform residual analysis on m4_lm again:



As residual constant variance got much more worse Deleting the Pricelog variable and re-adding the original Price variable to the data and rebuilding the model again with original Price variable:

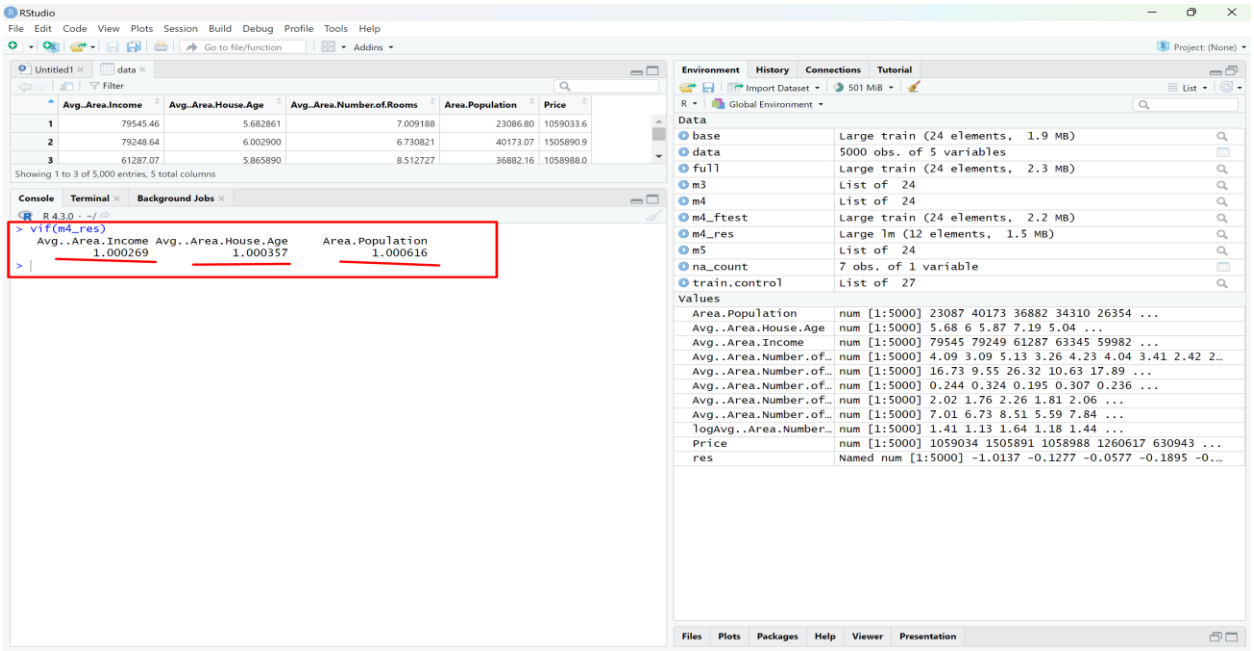


Qqplot:



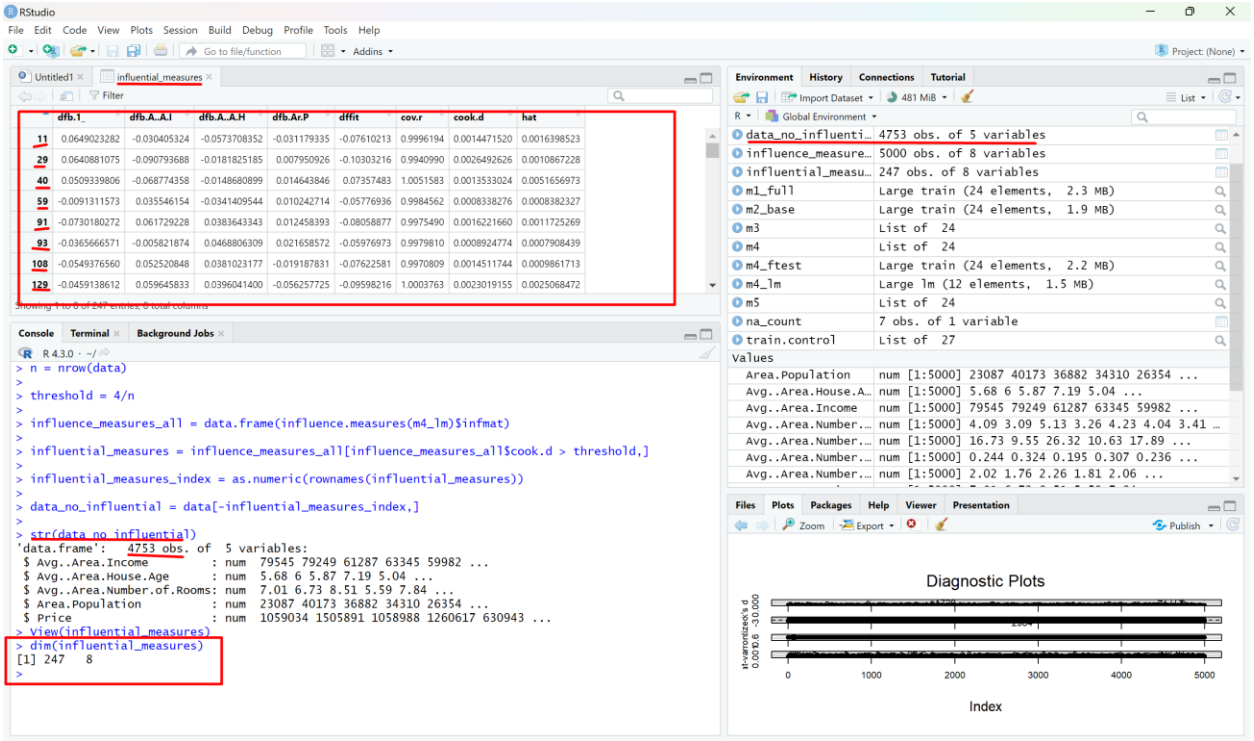
8. Checking multicollinearity for the selected model above (M4) by calculating VIF:

VIF for M4:

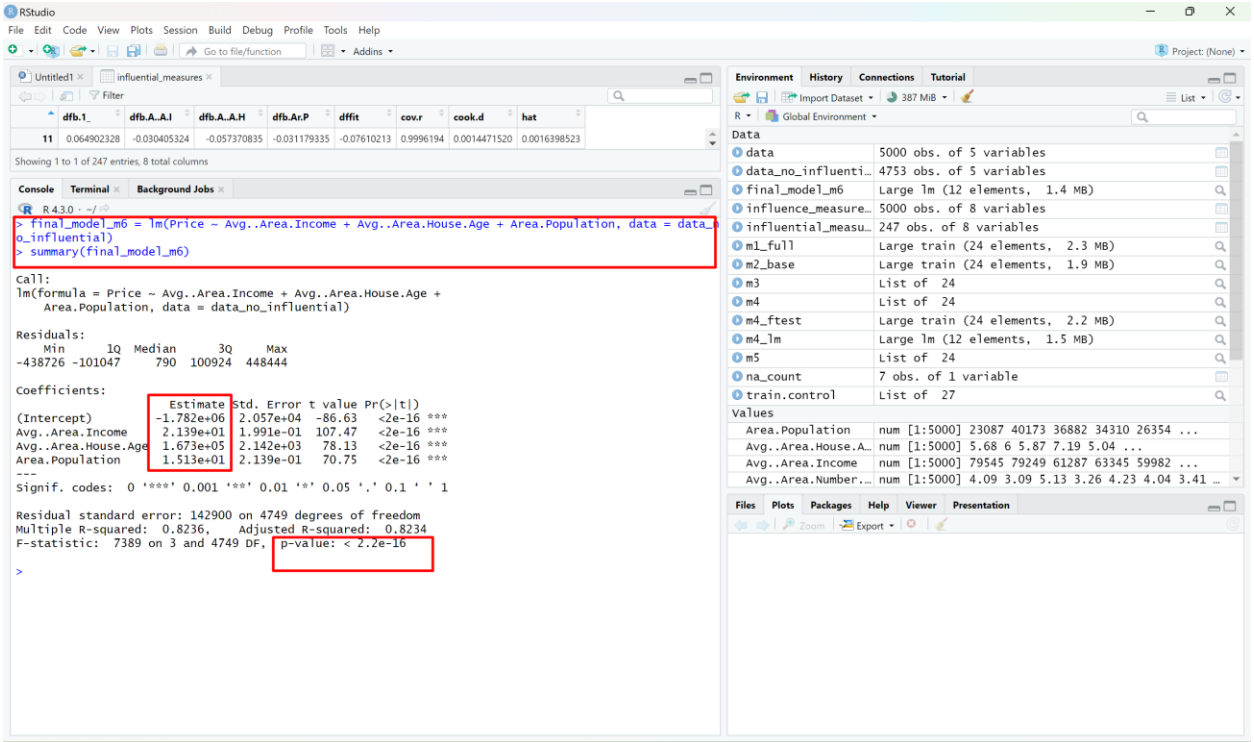


9. Checking for Influential points by cooks.distance

247 influential points were observed and removed from the data and created new data called data_no_influential



Rebuild the m4 again with the name “final model” with the new data (data_no_influential_points):



RMSE for the final model M6: