MTH 325, Quiz #6, Spring 2023 Name _____

Instructions: Answer each question as thoroughly as possible. Round answers to 4 decimal places as needed. Exact answers are best when possible. Be sure to answer all parts of each question.

1. Consider the data on home prices in **325quiz6data.xlsx**. Perform a natural log transformation to both variables and add them to the dataset. Then use best subset selection methods to find the best model to predict price from the other variables (excluding Home). Perform appropriate diagnostics and do the following:
   a. State your final equation (clearly state which variable is which)

The best 4-variable model is:

```
 (Intercept)    Bathrooms     Lot_Size        logHS        logLS
-301965.758    11657.417     5919.250    61647.095     5112.222
```

$$Price = 11{,}657.42Bathrooms + 5919.25LotSize + 61{,}647.10\ln(HomeSize) + 5112.22\ln(LotSize) - 301{,}965.76$$

Lowest adjR^2 and AIC

The best 3-variable model is:

```
(Intercept)    Bathrooms     Lot_Size        logHS
-316940.578    12804.263     7369.968    62810.606
```

$$Price = 12{,}804.26Bathrooms + 7369.97LotSize + 62{,}810.61\ln(HomeSize) - 316{,}940.58$$

Lowest BIC

   b. Conduct appropriate hypothesis tests on your final model for all coefficients.

Summary of 4-variable model:

```
Call:
lm(formula = Price ~ Bathrooms + Lot_Size + logHS + logLS, data = data6)

Residuals:
   Min     1Q Median     3Q    Max
-65169 -13344    273  12014  82148

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -301966      61865  -4.881 2.75e-06 ***
Bathrooms      11657       4299   2.712   0.0075 **
Lot_Size        5919       1307   4.529 1.23e-05 ***
logHS          61647       9027   6.829 2.18e-10 ***
logLS           5112       3483   1.468   0.1443
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23050 on 145 degrees of freedom
Multiple R-squared:  0.7014,   Adjusted R-squared:  0.6932
F-statistic: 85.15 on 4 and 145 DF,  p-value: < 2.2e-16
```

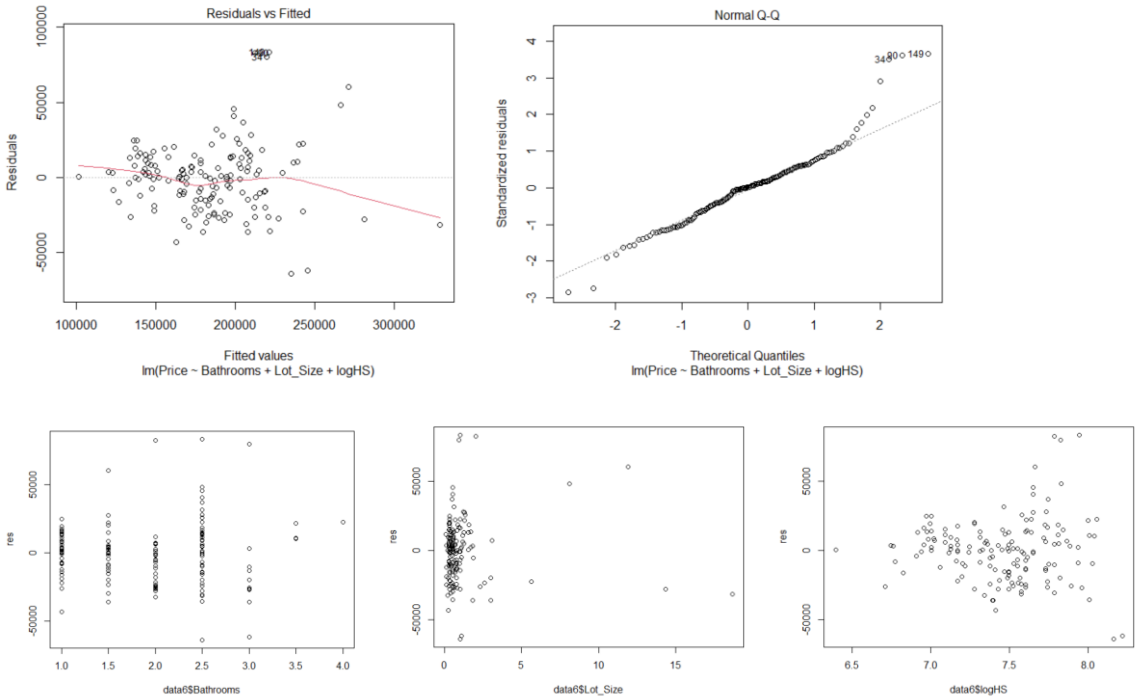
Summary of 3-variable model:

```
Call:
lm(formula = Price ~ Bathrooms + Lot_Size + logHS, data = data6)

Residuals:
   Min     1Q Median     3Q    Max
-64007 -14052    562  11525  83562
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -316940.6    61258.1  -5.174 7.45e-07 ***
Bathrooms     12804.3     4244.0   3.017  0.00301 **
Lot_Size       7370.0      858.3   8.587 1.23e-14 ***
logHS         62810.6     9028.0   6.957 1.08e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23140 on 146 degrees of freedom
Multiple R-squared:  0.697,    Adjusted R-squared:  0.6907
F-statistic: 111.9 on 3 and 146 DF,  p-value: < 2.2e-16
```

Based on the hypothesis tests, the 4-variable model is rejected since the 4th variable is not significant. Therefore, we use the 3-variable model as the final model since all the variables here are significant.

c. Create residual plots and analyze them to test your assumptions for the multivariable model.

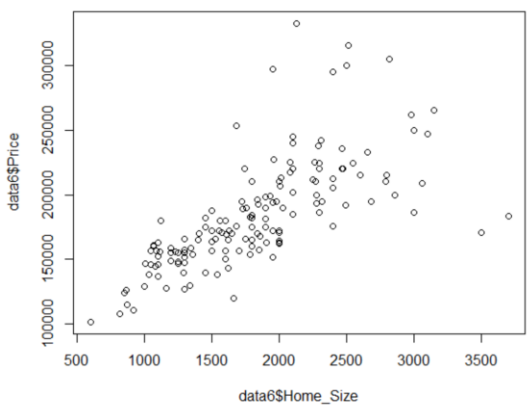


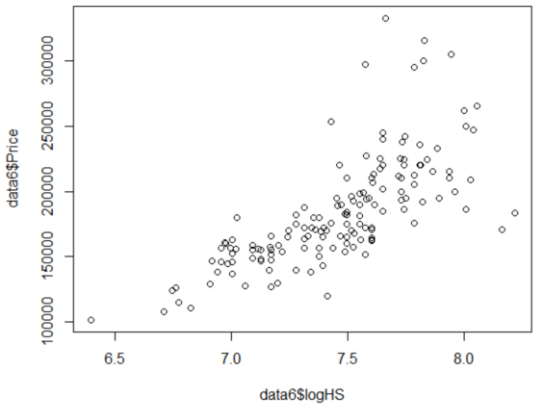
Bathrooms and LotSize look okay. The logHS variable still looks like it’s maybe heteroscedastic.

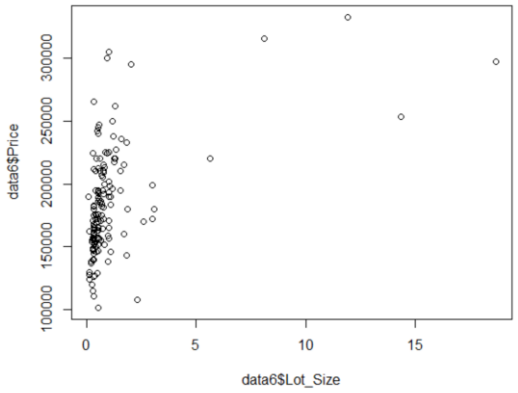
d. What is the final R^2 of the model? What does it mean in context?

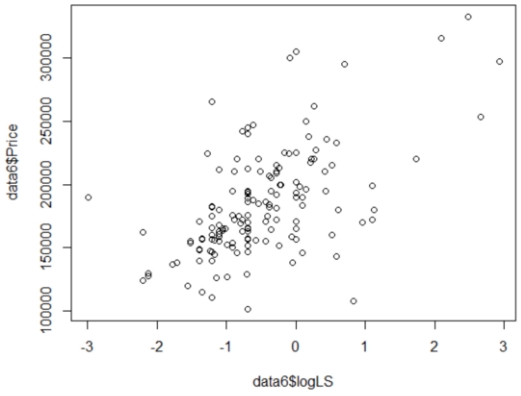
The final R^2 value is 0.697 so about 69.7% of the variability in price can be explained by the number of bathrooms, the lot size and the log of the Home Size.

e. Create a scatter plot of the transformed variables relative to the price, and the untransformed variables relative to the price. Does the transformation appeared to have improved the linearity? Explain.









Not much change from Home Size to log of Home Size. The Lot Size is much more spread out and does look better. But, in the end, it's the regular Lot Size and survives in the final model (maybe contributing to the heteroscedacity).