Multiple linear regression

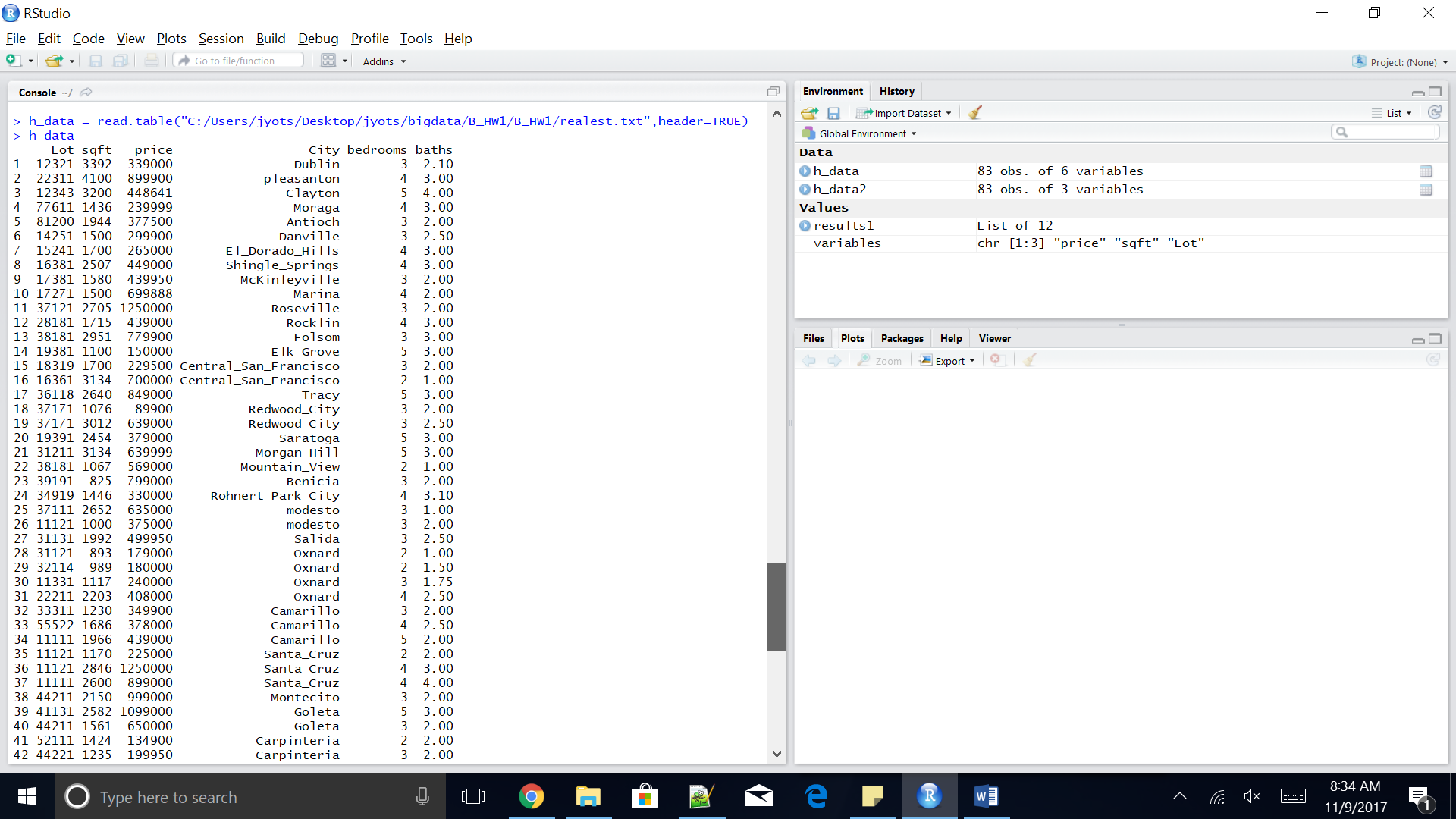
Multiple linear regression analysis using data collected for approximate 100 real estate properties (data attached realest.txt). Will analyze which variable predicts better the price of these properties.

Steps to perform:

1: Load the data to Rstudio

h\_data = read.table("C:/Users/jyots/Desktop/jyots/bigdata/B\_HW1/B\_HW1/realest.txt",header=TRUE)

h\_data



2: Plot multi variant regression chart for variables Price, Size and Lot

variables = c("price", "sqft", "Lot")

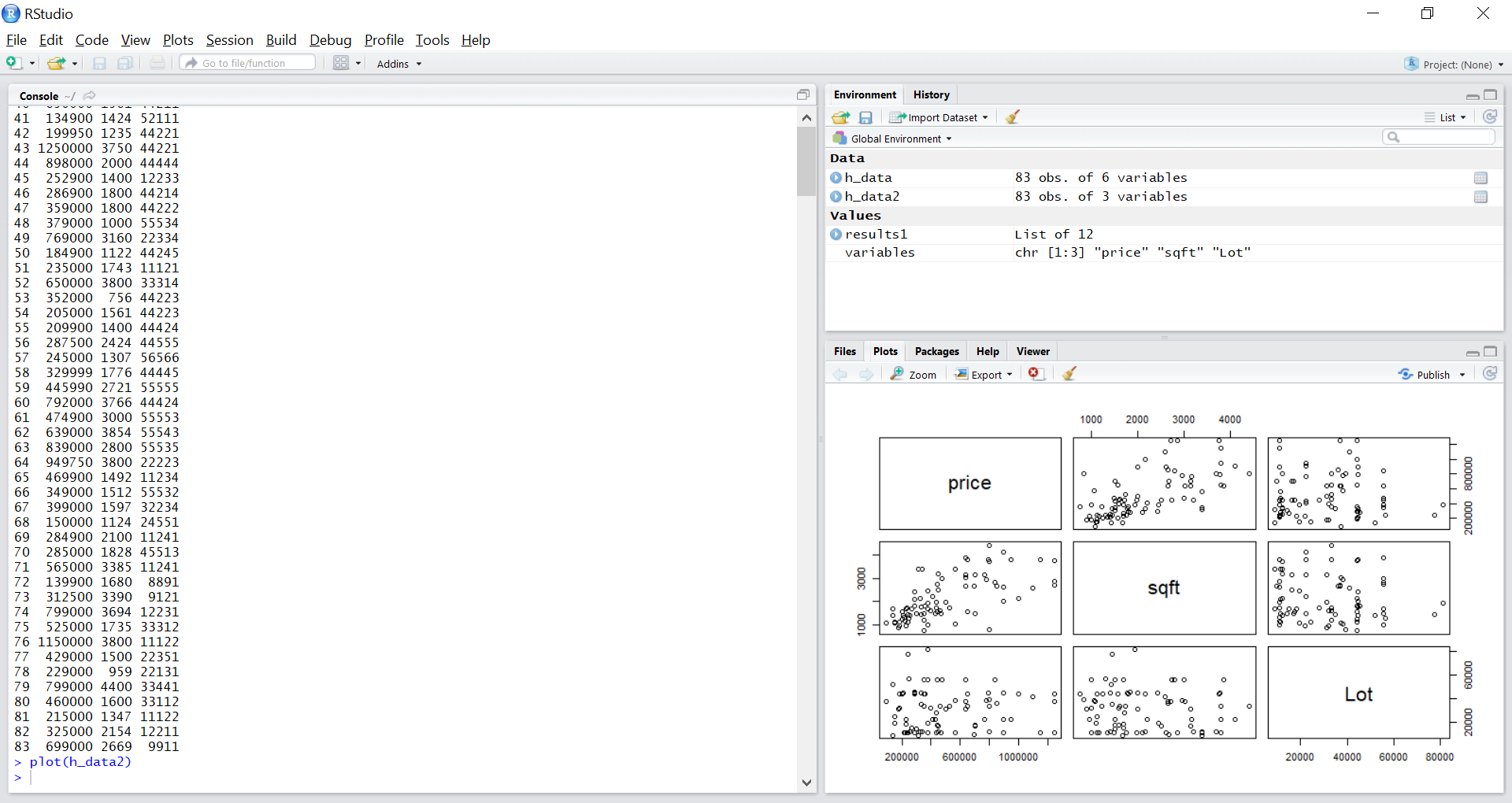
h\_data2 = h\_data[variables]

h\_data2

plot(h\_data2)

results1 = lm(price ~ sqft + Lot, data=h\_data)

results1



3: Calculate the multivariable linear model using lm function for above variables and calculate the intercept and parameters “A1” and “A2” for the following function:

Y= Intercept +A1\*X1 +A2\*X2

Where Y= Price, X1,X2 are Size and Lot variables. A1 and A2 are regression coefficients.

Assume that these two variables have no effect on the price of the properties.

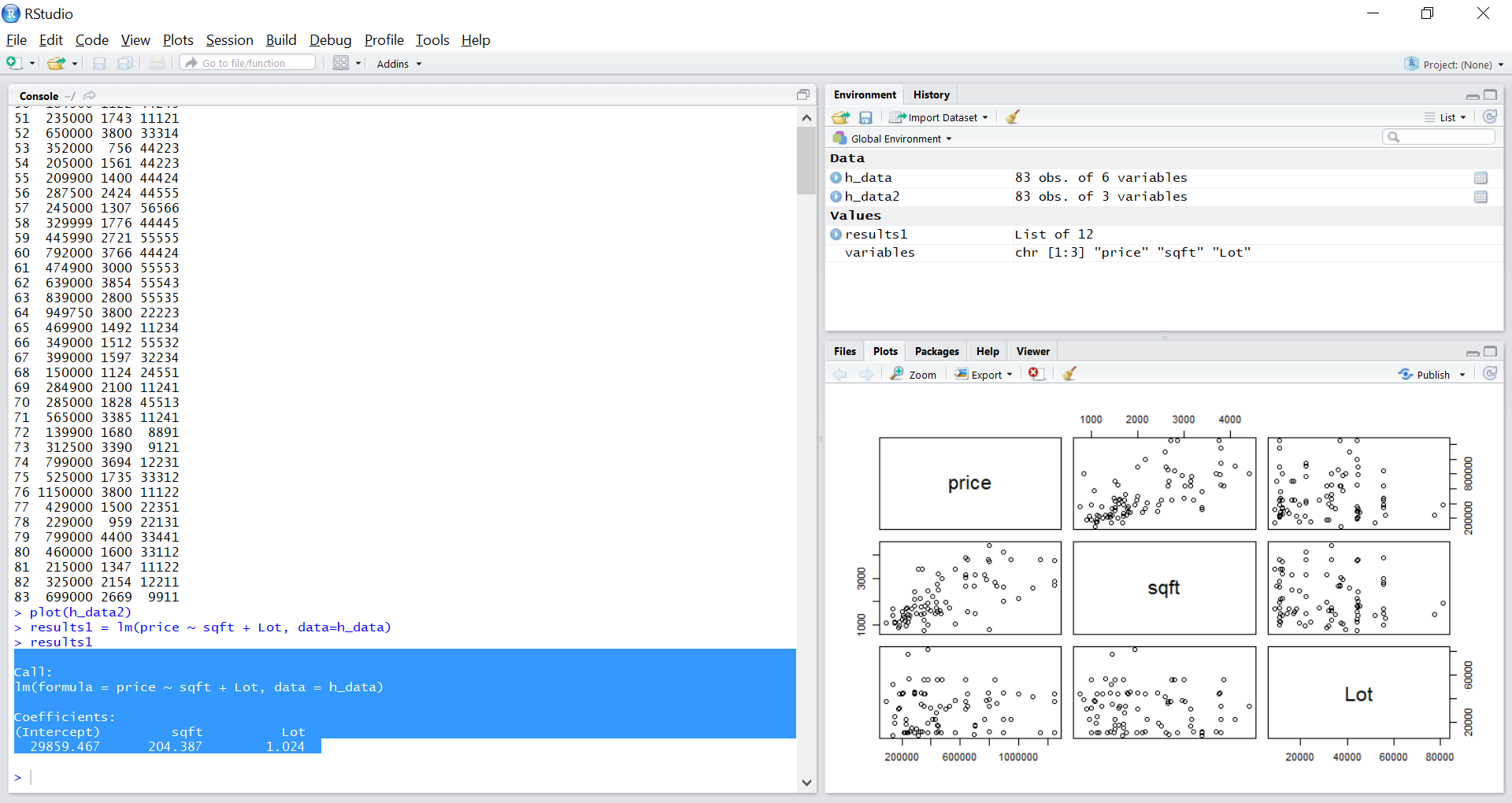
Call:

lm(formula = price ~ sqft + Lot, data = h\_data)

Coefficients:

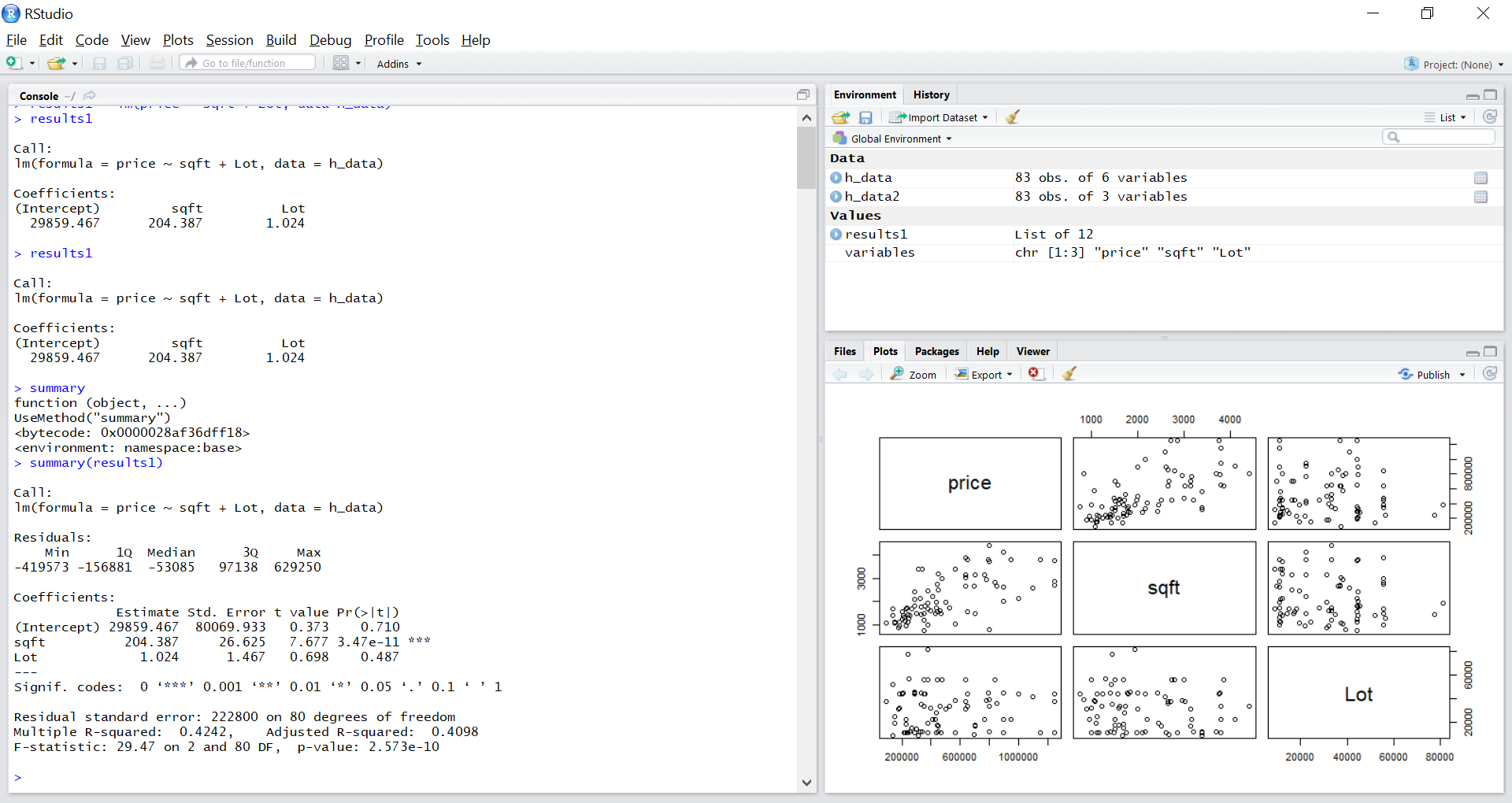
(Intercept) sqft Lot

29859.467 204.387 1.024



4: Calculate the summary for the “lm” linearized value and find the distribution for Size and Lot, their P values and the overall F statistic and P value for this regression function.

Interpret the results and confirm or disconfirm the initial hypothesis that variables Size and Lot are not predicting the cost of the properties.



Summary

Call:

lm(formula = price ~ sqft + Lot, data = h\_data)

Residuals:

Min 1Q Median 3Q Max

-419573 -156881 -53085 97138 629250

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 29859.467 80069.933 0.373 0.710

sqft 204.387 26.625 7.677 3.47e-11 \*\*\*

Lot 1.024 1.467 0.698 0.487

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 222800 on 80 degrees of freedom

Multiple R-squared: 0.4242, Adjusted R-squared: 0.4098

F-statistic: 29.47 on 2 and 80 DF, p-value: 2.573e-10

In the output F = 29.47 and p : 2.573e-10

which is less than 0.05 or 0.01 we will

reject the null hypothesis and suggest variables Size and Lot have no

effect on Price.

Lot (p = 0.487) which is greater than the signuficance level of 0.05

sqft (p=3.47e-11).

5: Repeat the same multi variable linear regression analyzes, except now include two additional variables, such as number of bathroom and bedroom (include additional two variables). Establish hypothesis 2 about whether the # of bathrooms and bedrooms are predicting the value of these realestates.

variables2 = c("price", "bedrooms", "baths")

h\_data3 = h\_data[variables2]

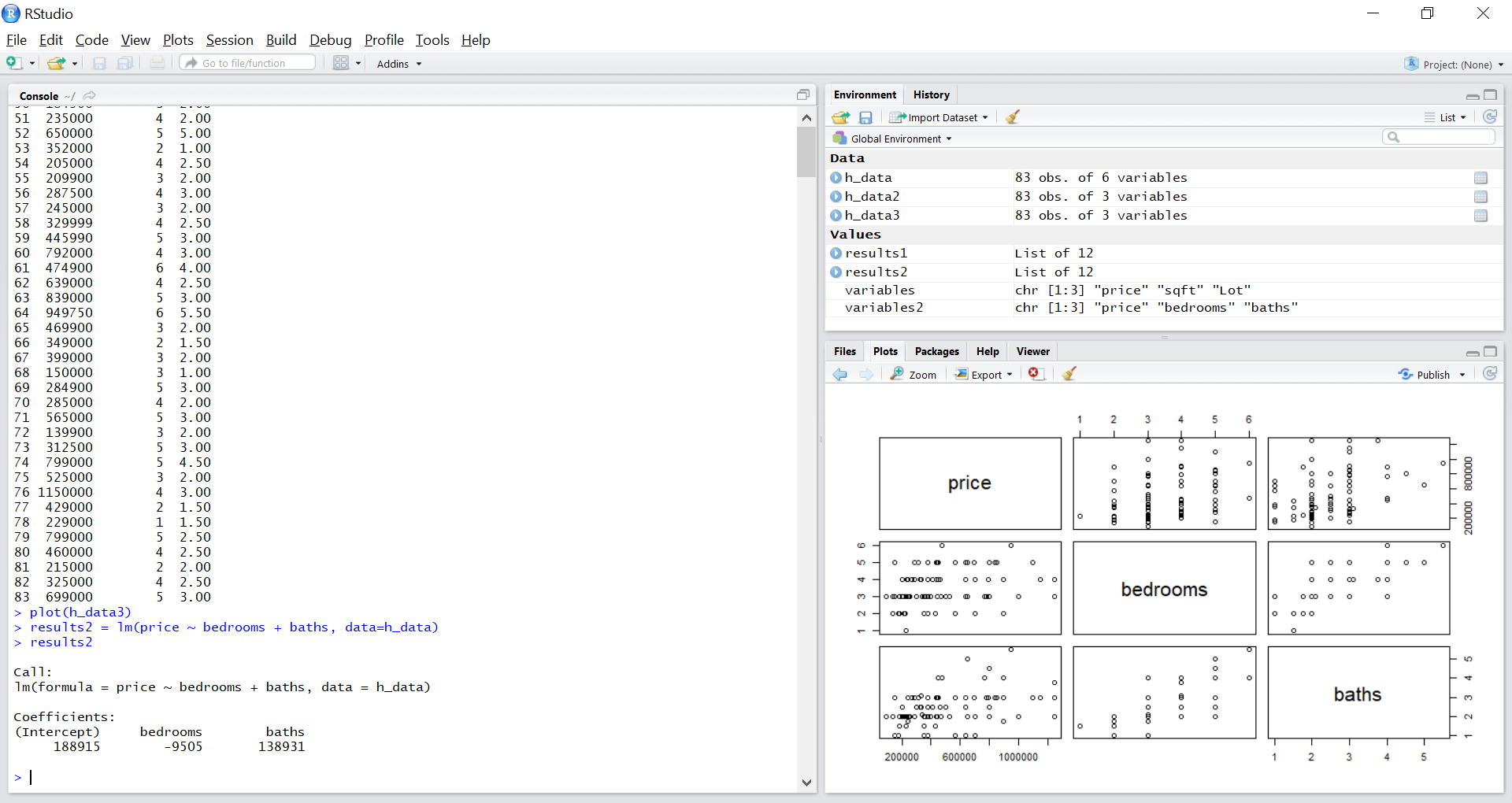
h\_data3

plot(h\_data3)

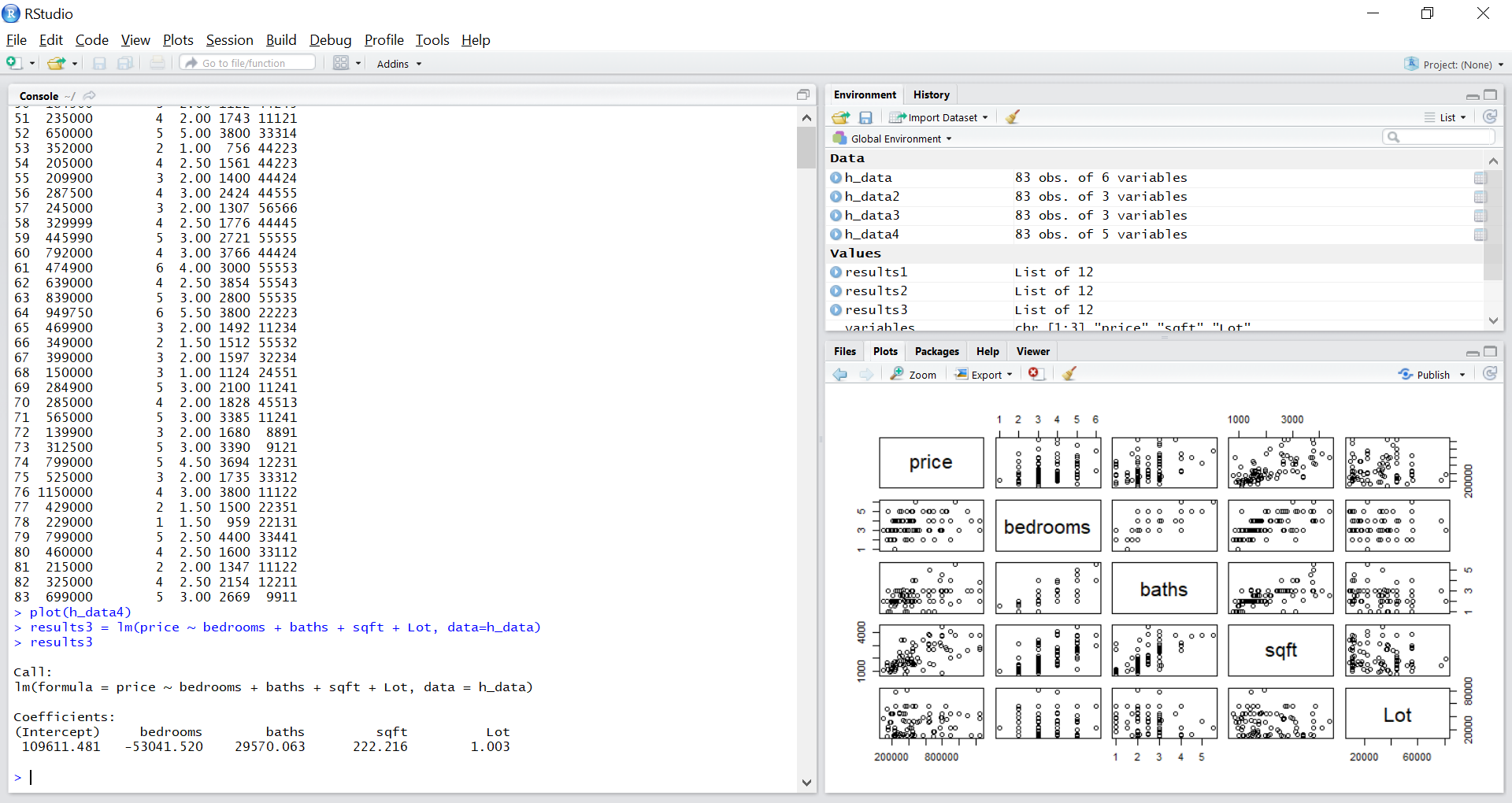
results2 = lm(price ~ bedrooms + baths, data=h\_data)

results2

5.1. Create the multi-variant plot chart and interpret what you see



5.2. Generate the linear fitting model (lm function) for the 4 variables: Size, Lot, # bathrooms, #bedrooms. Provide the expanded Y function.



Call:

lm(formula = price ~ bedrooms + baths + sqft + Lot, data = h\_data)

Coefficients:

(Intercept) bedrooms baths sqft Lot

109611.481 -53041.520 29570.063 222.216 1.003

5.3. Calculate the summary distribution model for he obtained lm model and interpret the results. Confirm or disconfirm the 2nd hypothesis that the # of bathrooms and # of bedrooms are not predicting the price of the properties. Interpret the results.

summary(results3)

Call:

lm(formula = price ~ bedrooms + baths + sqft + Lot, data = h\_data)

Residuals:

Min 1Q Median 3Q Max

-439699 -127305 -75190 91925 620263

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 109611.481 101820.971 1.077 0.285

bedrooms -53041.520 35878.134 -1.478 0.143

baths 29570.063 46629.381 0.634 0.528

sqft 222.216 35.332 6.289 1.71e-08 \*\*\*

Lot 1.003 1.468 0.683 0.496

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 222400 on 78 degrees of freedom

Multiple R-squared: 0.4408, Adjusted R-squared: 0.4121

F-statistic: 15.37 on 4 and 78 DF, p-value: 2.608e-09

In the output F = 15.37 and p : 2.608e-09

which is less than 0.05 or 0.01 we will

reject the null hypothesis and suggest variables have

effect on Price.

bedrooms (p = 0.285) which is greater than the significance level of 0.05

baths (p=0.528). which is greater than the significance level of 0.05

6: Perform ANOVA calculation for the initial lm() model (call it “reduced model”) and the second expanded model (call it “full model”).

Interpret the resulting F test and P value and confirm/disconfirm whether the reduced model predicts sufficiently the outcome and whether the additional two variables (# of bedrooms and bathrooms) have or don’t have sufficient contribution to the full model. Explain whether their contributions can be considered non-significant or not.

ANSWER:

Analysis of Variance Table

Model 1: price ~ sqft + Lot

Model 2: price ~ bedrooms + baths + sqft + Lot

Res.Df RSS Df Sum of Sq F Pr(>F)

1 80 3.9727e+12

2 78 3.8586e+12 2 1.1412e+11 1.1534 0.3209

Steps are written below:

h\_data = read.table("C:/Users/jyots/Desktop/jyots/bigdata/B\_HW1/B\_HW1/realest.txt",header=TRUE)

h\_data

variables = c("price", "sqft", "Lot")

h\_data2 = h\_data[variables]

h\_data2

plot(h\_data2)

results1 = lm(price ~ sqft + Lot, data=h\_data)

results1

variables2 = c("price", "bedrooms", "baths")

h\_data3 = h\_data[variables2]

h\_data3

plot(h\_data3)

results2 = lm(price ~ bedrooms + baths, data=h\_data)

results2

variables3 = c("price", "bedrooms", "baths","sqft", "Lot")

h\_data4 = h\_data[variables3]

h\_data4

plot(h\_data4)

results3 = lm(price ~ bedrooms + baths + sqft + Lot, data=h\_data)

results3

aReduce\_model = lm(price ~ sqft + Lot, data=h\_data)

Full\_model = lm(price ~ sqft + Lot + bathrooms + bedrooms, data= h\_data)

anova(results1, results3)

*Some guidance here to help:*

*Reduce\_model = lm(price ~ Size + Lot, data=realest)*

*Full\_model = lm(price ~ Size + Lot + bathrooms + bedrooms, data= realest)*

*anova(reduced\_model, full\_model)*

**STEPS***:*

h\_data = read.table("C:/Users/jyots/Desktop/jyots/bigdata/B\_HW1/B\_HW1/realest.txt",header=TRUE)

h\_data

variables = c("price", "sqft", "Lot")

h\_data2 = h\_data[variables]

h\_data2

plot(h\_data2)

results1 = lm(price ~ sqft + Lot, data=h\_data)

results1

variables2 = c("price", "bedrooms", "baths")

h\_data3 = h\_data[variables2]

h\_data3

plot(h\_data3)

results2 = lm(price ~ bedrooms + baths, data=h\_data)

results2

variables3 = c("price", "bedrooms", "baths","sqft", "Lot")

h\_data4 = h\_data[variables3]

h\_data4

plot(h\_data4)

results3 = lm(price ~ bedrooms + baths + sqft + Lot, data=h\_data)

results3

aReduce\_model = lm(price ~ sqft + Lot, data=h\_data)

Full\_model = lm(price ~ sqft + Lot + bathrooms + bedrooms, data= h\_data)

anova(results1, results3)