# Lesson 19 SLO Real Estate (5.2)

LTC J.K. Starling

**Background** Although numerous variables impact the initial selling price of a home, realtors try to identify which variables are most critical in a particular region. Using information on similar homes helps them establish a fair market value for a new home. In this exploration, you will explore home prices from a particular housing market to help to help identify the impact of several variables on the list price.

```
# Load packages
library(tidyverse)
library(ggResidpanel)
library(car)
library(GGally)
```

**STEP 1: Ask a research question.**

(1) Suggest some potential characteristics (variables) of homes that are easily accessible and that you believe will help explain variation in home prices.

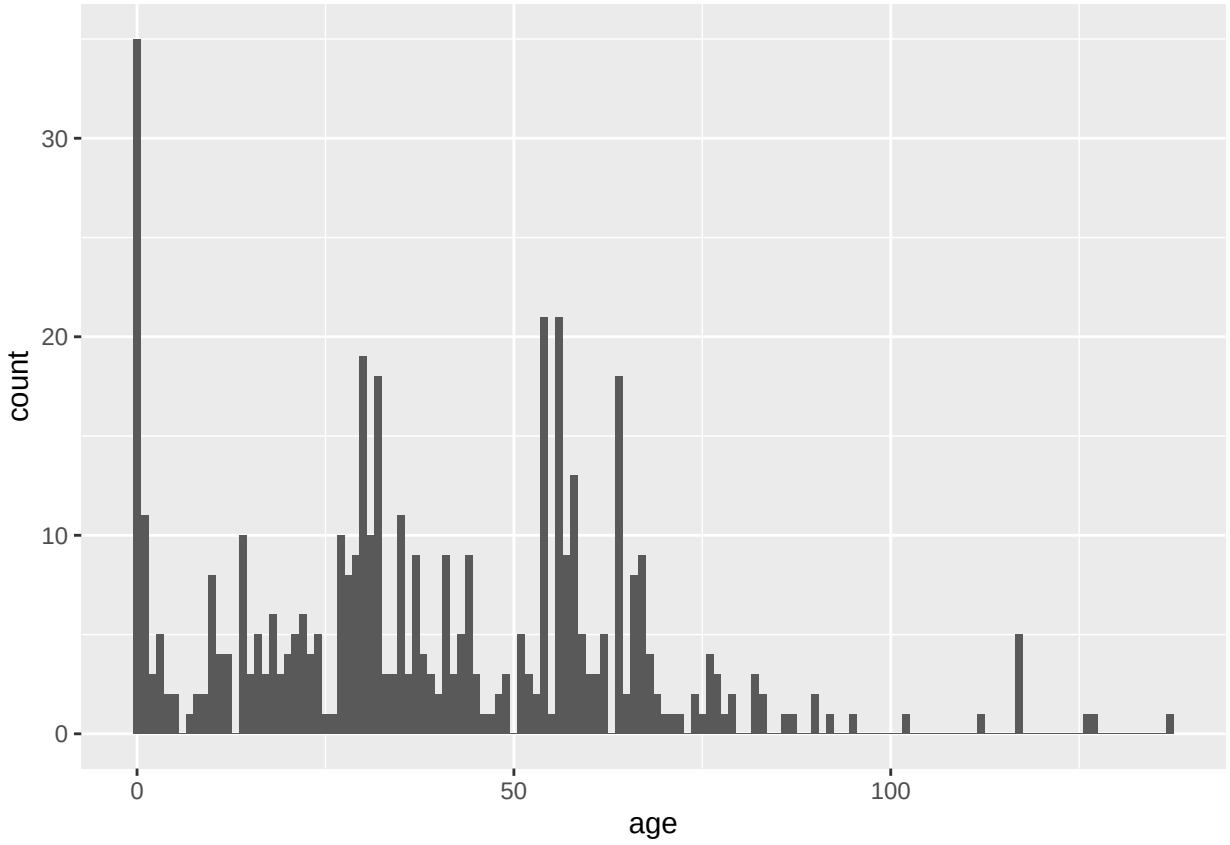
(2) Suppose you thought both the size of the home and the age of the home were important variables. How might you decide which one was more important?

**STEP 2: Design a study and collect data.** A local realtor shared data with us on all listings in San Luis Obispo, CA, in 2017. We did some data cleaning (e.g., a home listed as 4200 square feet built in 1910 was excluded since online follow-up on that home could not confirm that interesting combination of values). A few other homes with some missing data were also removed. The data in SLO2017 contains list prices (in $100,000s) for 454 homes. Some quantitative variables available for predicting the list price include square footage (in thousands of square feet), number of bedrooms, and year built.

```
### Data management
SLO <- read.csv("https://raw.githubusercontent.com/jkstarling/MA376/main/SLO2017.csv", header = TRUE, s
# setwd("C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons/Block III/LSN19/")
# SLO <- read.csv("SLO2017.csv", header = TRUE, stringsAsFactors = TRUE)
```

(3) Create a new variable that represents the age of the home in 2017. Show a histogram of these values. Do these values make sense?

```
SLO <- SLO %>% mutate(age = 2017 - YrBuilt)
SLO %>% ggplot(aes(x = age)) + geom_histogram(binwidth = 1)
```




```
max(SLO$YrBuilt)
```

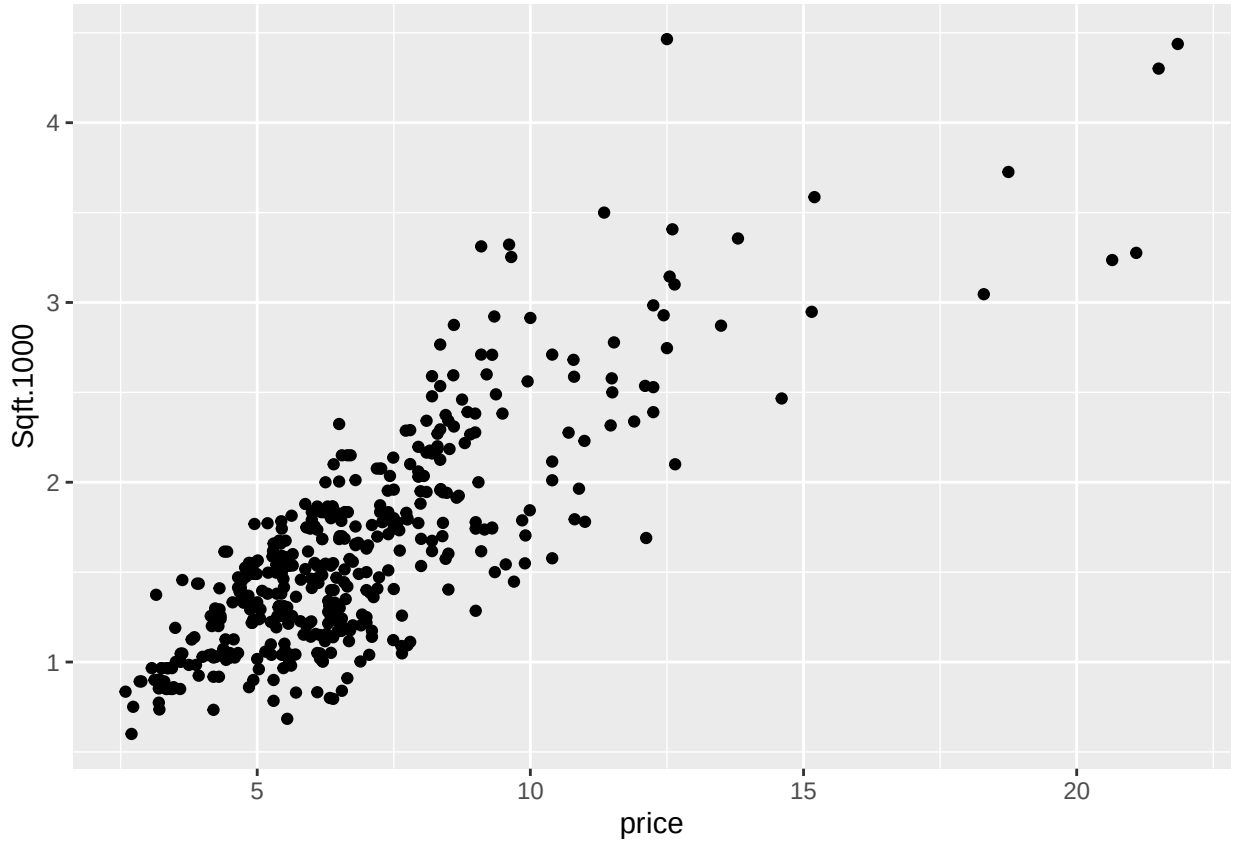
```
## [1] 2017
```

**STEP 3: Explore the data.**

(4) Examine scatterplots and summarize how each of these variables (sqft, number of bedrooms, and age) are associated with list price and how the pairs of explanatory variables are associated with each other.

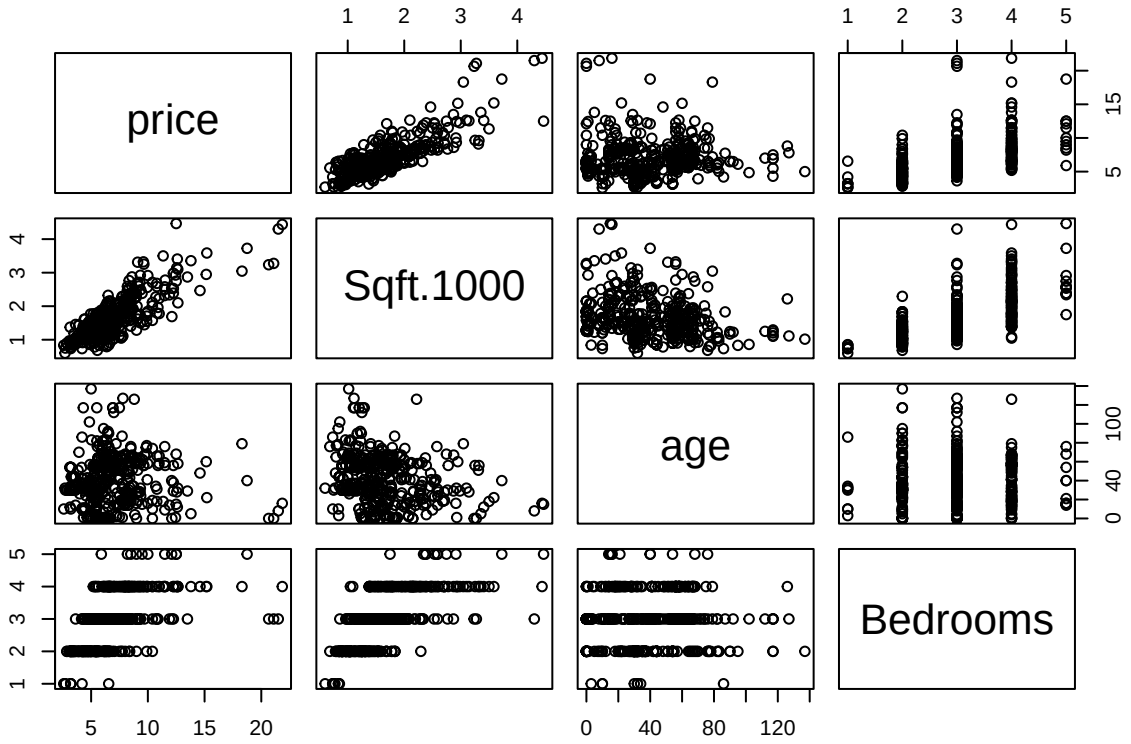
- Price vs. square footage:
- Price vs. number of bedrooms:
- Price vs. age:
- Square footage vs. age:
- Square footage vs. number of bedrooms:
- Age vs. number of bedrooms:

```
SLO %>% ggplot(aes(x=price, y = Sqft.1000)) + geom_point()
```

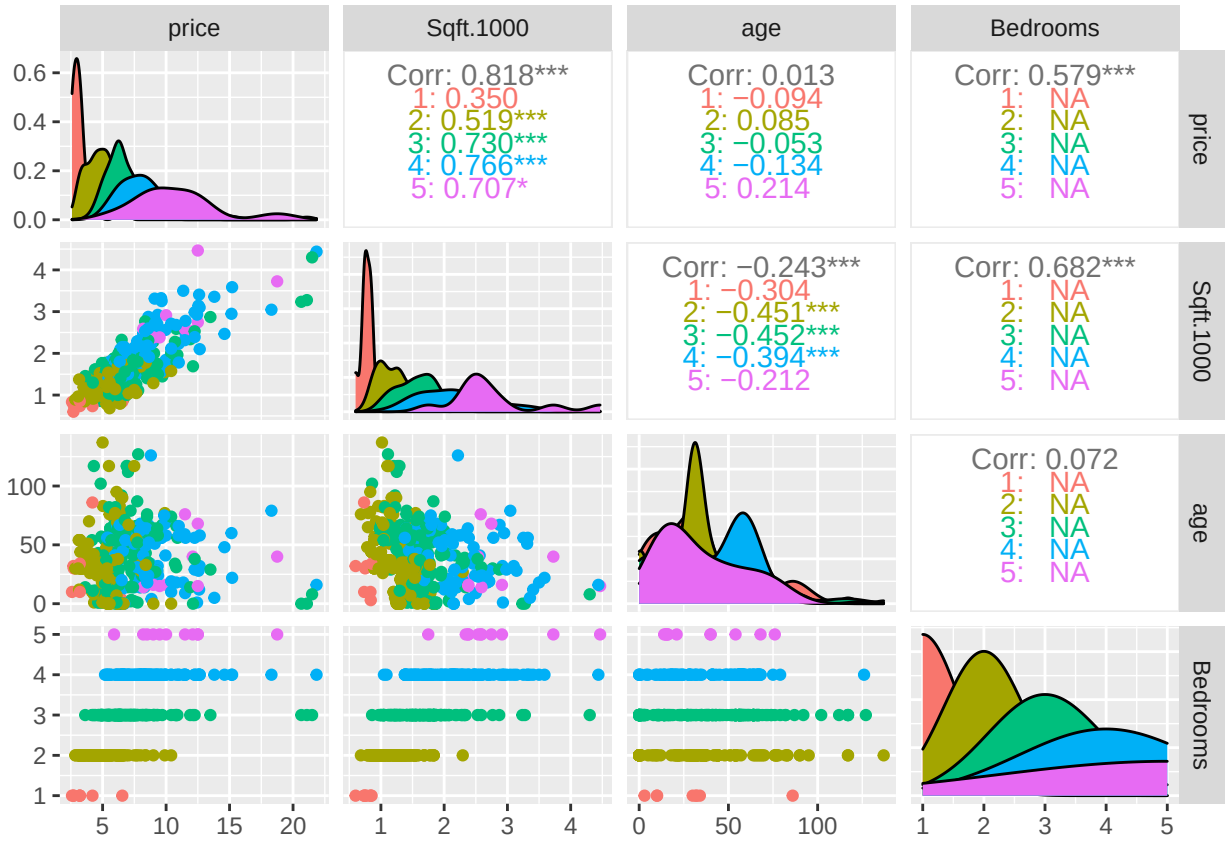



```
# SLO %>% ggplot(aes(x=price, y = Bedrooms)) + geom_point()
# SLO %>% ggplot(aes(x=price, y = age)) + geom_point()
# SLO %>% ggplot(aes(x=Sqft.1000, y = age)) + geom_point()
# SLO %>% ggplot(aes(x=Sqft.1000, y = Bedrooms)) + geom_point()
# SLO %>% ggplot(aes(x=age, y = Bedrooms)) + geom_point()

SLO2 <- SLO %>% select(price, Sqft.1000, age, Bedrooms)
pairs(SLO2)
```




```
# ggpairs(SL02)
ggpairs(SL02, aes(color = as.factor(Bedrooms)))
```




(5) How accurately can we predict the list price based only on the *square footage of the home*?

```
sq.lm <- SLO2 %>% lm(price ~ Sqft.1000, data = .)
summary(sq.lm)
```

```
##
## Call:
## lm(formula = price ~ Sqft.1000, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7178 -1.1129 -0.2978  0.8520  8.2311
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.8488     0.2100   4.042 6.22e-05 ***
## Sqft.1000     3.6661     0.1214  30.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.584 on 452 degrees of freedom
## Multiple R-squared:  0.6685, Adjusted R-squared:  0.6678
```

```
## F-statistic: 911.5 on 1 and 452 DF,  p-value: < 2.2e-16
```

(6) How accurately can we predict the list price based only on the *age of the home*?

```
age.lm <- SLO2 %>% lm(price ~ age, data = .)
summary(age.lm)
```

```
##
## Call:
## lm(formula = price ~ age, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1482 -1.6536 -0.4702  1.1188 15.1036
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.724567   0.234644  28.659   <2e-16 ***
## age         0.001363   0.005003   0.272    0.785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.752 on 452 degrees of freedom
## Multiple R-squared:  0.0001641,  Adjusted R-squared:  -0.002048
## F-statistic: 0.0742 on 1 and 452 DF,  p-value: 0.7854
```

**STEP 4: Draw inferences beyond the data.**

(7) Now explore the two-variable model that includes both Sqft/1000 and age.

```
sfage.lm <- SLO2 %>% lm(price~Sqft.1000+age, data = .)
summary(sfage.lm)
```

```
##
## Call:
## lm(formula = price ~ Sqft.1000 + age, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8381 -0.8109 -0.1915  0.6727  8.7612
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.484322   0.247717  -1.955   0.0512 .
```

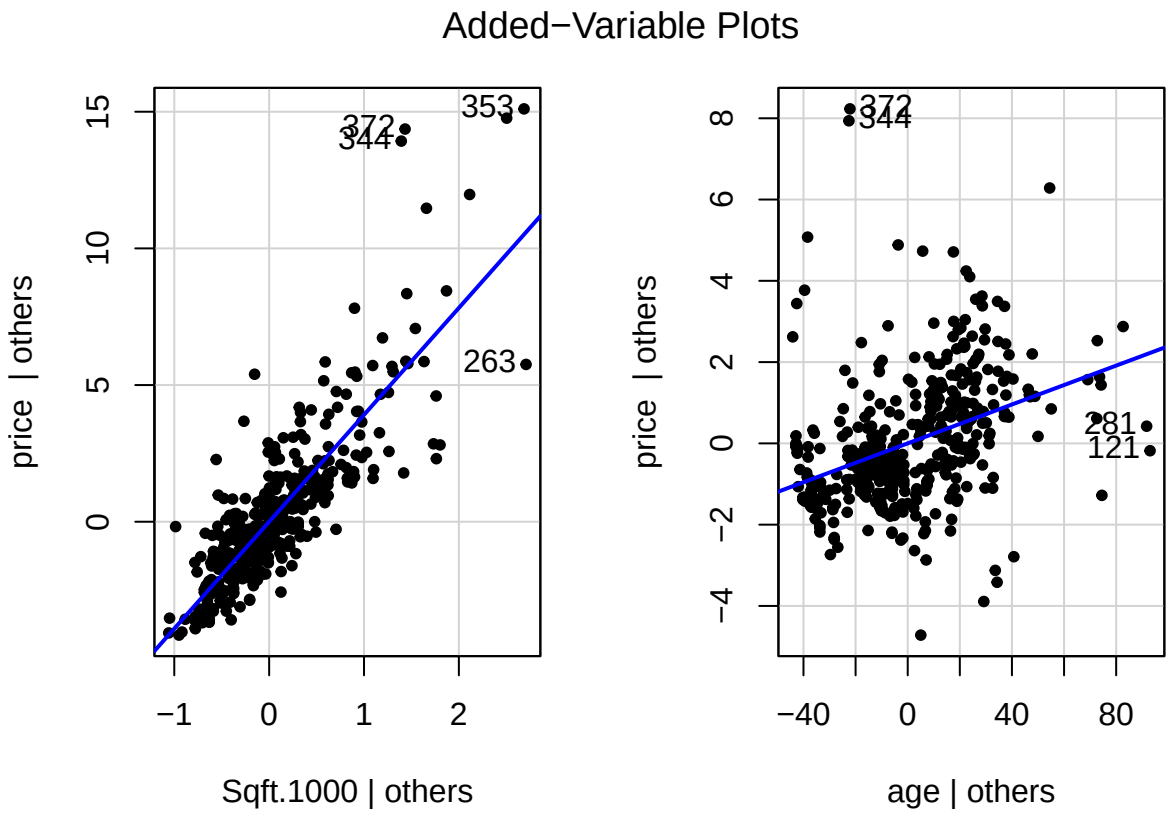
```
## Sqft.1000    3.911209   0.115979  33.724   <2e-16 ***
## age          0.023924   0.002752   8.694   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.468 on 451 degrees of freedom
## Multiple R-squared:  0.7161, Adjusted R-squared:  0.7148
## F-statistic: 568.8 on 2 and 451 DF,  p-value: < 2.2e-16
```

(8) Is the two-variable model statistically significant? Do the slope coefficients change? Do the variables still explain variation in the list prices after adjusting for the other variable? How does the model $R^2$ value compare to the sum of the individual $R^2$ values?

(9) Based on your analyses so far, are sqft and age confounded? Explain how you are deciding.

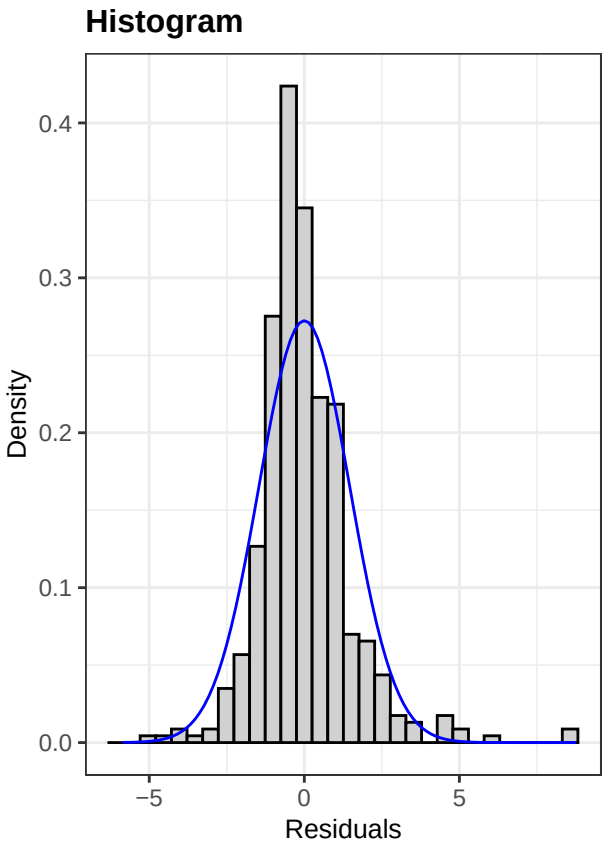
(10) Create added variable plots using the `avPlots()` function. Describe what you see (use the figure on the *right*).

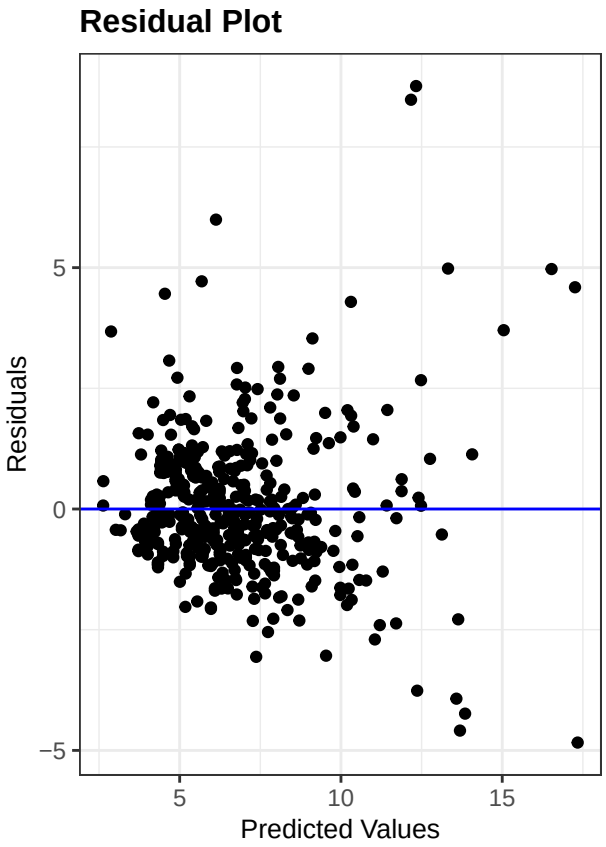
```
# from car package
avPlots(sfage.lm, pch=20)
```




(11) Based on the residual plots, does this statistical model appear to be valid?

```
resid_panel(sfage.lm, plots = c('hist', 'resid'))
```




(12) Next, we want to consider the potential interaction between age and square footage. Keep in mind this is very different from asking whether age and square footage are associated as you just explored. (note: in the SLO file there is already a column with the interaction term.)

Create a linear model that includes interaction between age and square footage. Is the interaction statistically significant (after adjusting for the other two variables)? Cite appropriate information to support your conclusion.

```
SLO3 <- SLO %>% select(price, Sqft.1000, age, Bedrooms, SqftxAge)
int.lm <- SLO3 %>% lm(price~Sqft.1000+age+SqftxAge, data=.)
int.lm <- SLO3 %>% lm(price~Sqft.1000*age, data=.)
summary(int.lm)
```

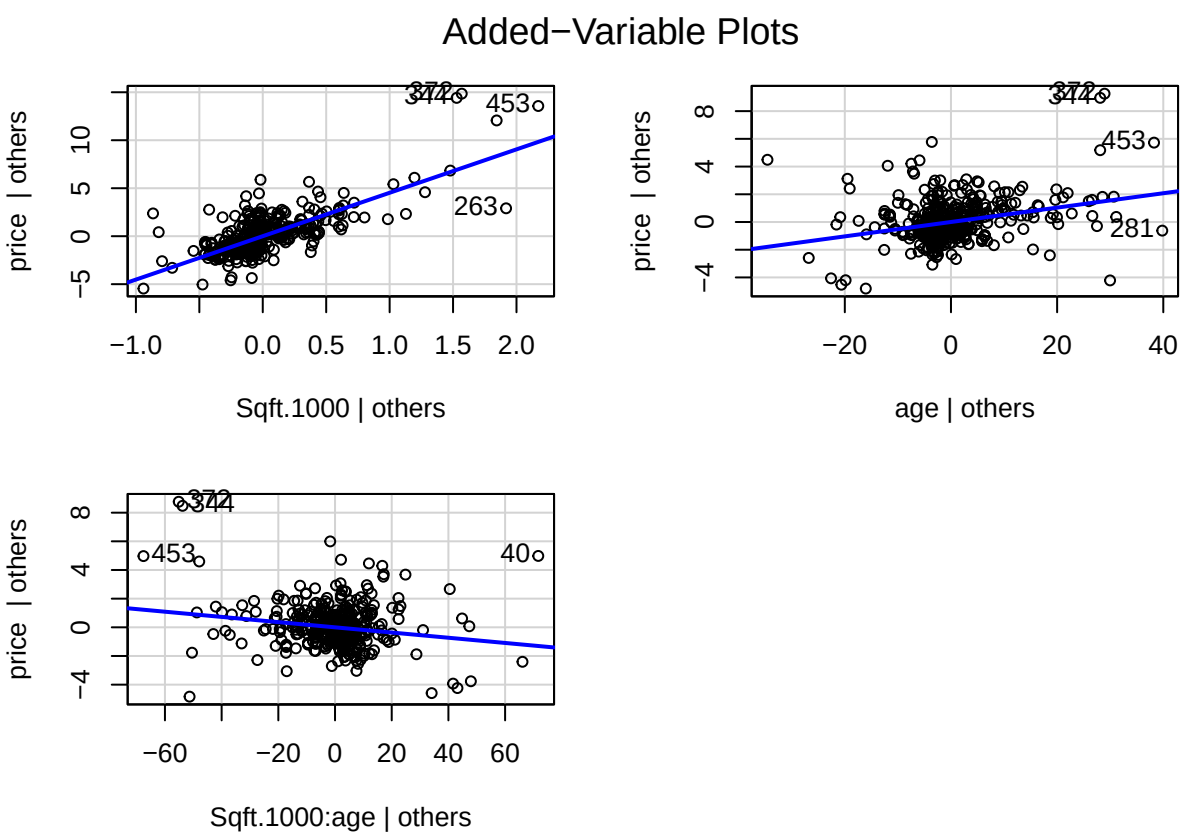
```
##
## Call:
## lm(formula = price ~ Sqft.1000 * age, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7703 -0.7512 -0.1991  0.5868  7.7595
##
## Coefficients:
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.488252   0.360263  -4.131 4.31e-05 ***
## Sqft.1000      4.523440   0.197909  22.856  < 2e-16 ***
## age            0.051874   0.007859   6.601 1.15e-10 ***
## Sqft.1000:age -0.018168   0.004794  -3.789 0.000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.447 on 450 degrees of freedom
## Multiple R-squared:  0.7249, Adjusted R-squared:  0.723
## F-statistic: 395.2 on 3 and 450 DF,  p-value: < 2.2e-16
```

```
anova(int.lm)
```

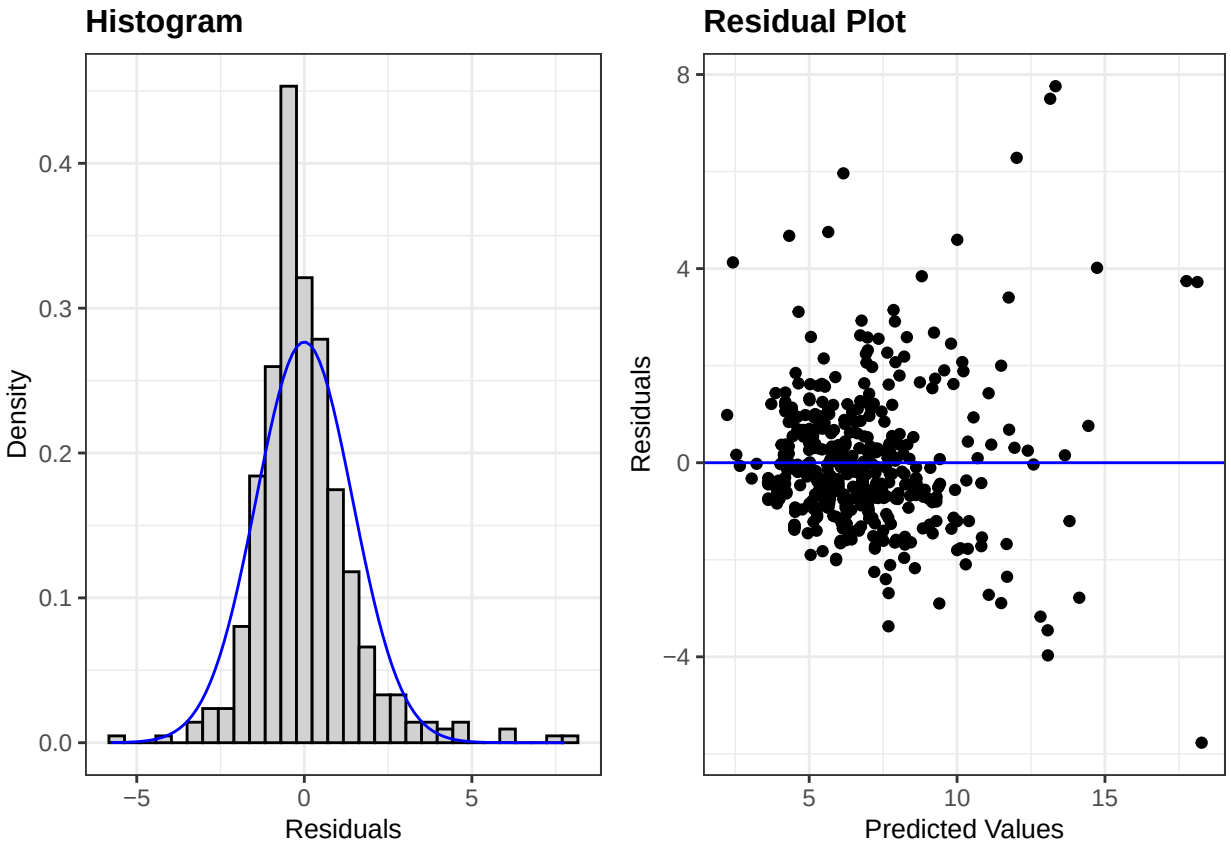
```
## Analysis of Variance Table
##
## Response: price
##                Df  Sum Sq Mean Sq  F value    Pr(>F)
## Sqft.1000       1 2288.48 2288.48 1093.407 < 2.2e-16 ***
## age             1  162.89  162.89   77.826 < 2.2e-16 ***
## Sqft.1000:age   1   30.05   30.05   14.359 0.0001716 ***
## Residuals     450  941.84    2.09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
avPlots(int.lm)
```



(13) Has the behavior of the residual plots improved with the inclusion of the interaction? Explain.

```
resid_panel(int.lm, plots = c('hist', 'resid'))
```




(14) Is the slope coefficient for the interaction positive or negative? What does the sign of the coefficient of the interaction tell you about the nature of the interaction in this context?

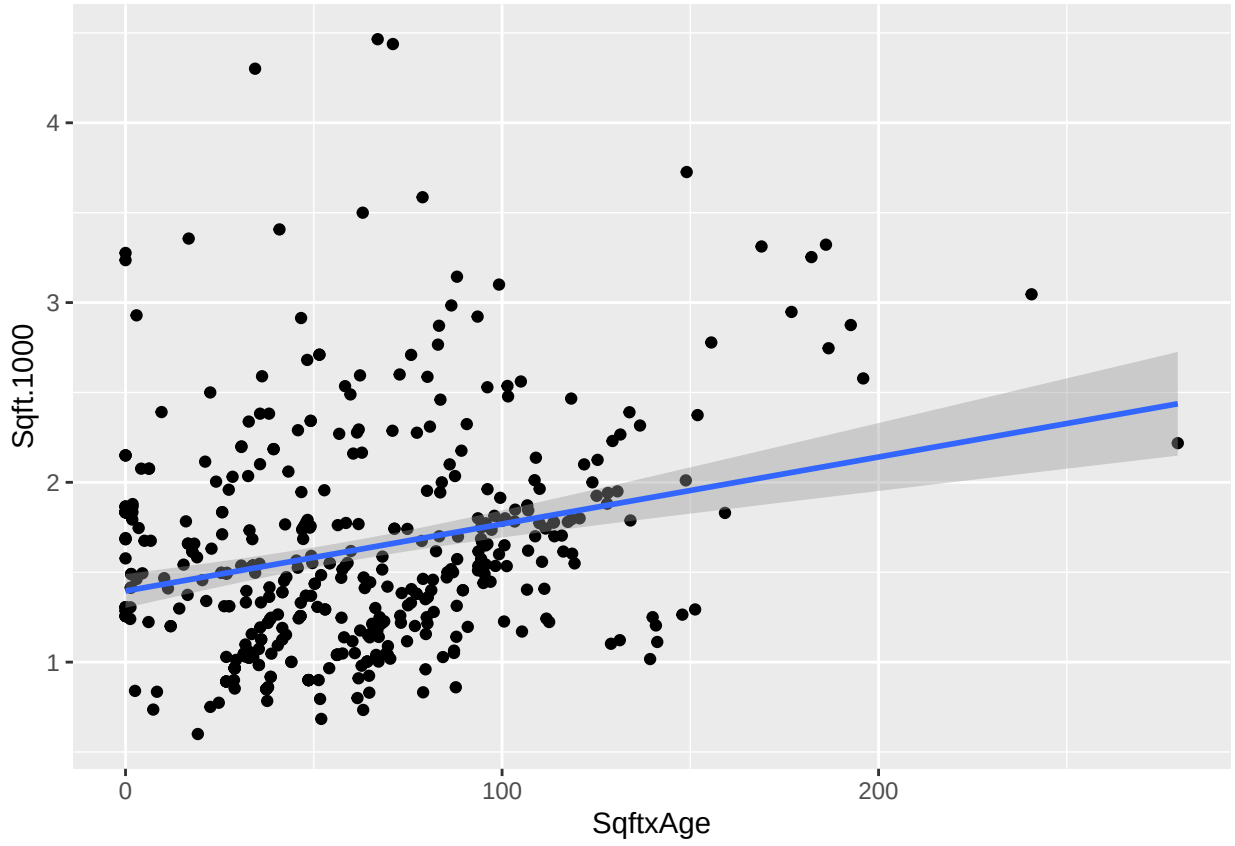
(15) To further illustrate your answer to #14, write out the prediction equations for predicting price from square footage for the following ages:

- 20:
- 70:
- 120:

(16) How did the "significance" of sqft and of age change when the interaction was included in the model? (Hint: Look at the t-statistics.)

(17) Create a scatterplot of the interaction term vs. square footage. Do you see evidence of a linear association between the interaction term and square footage? Why is this not a huge surprise?

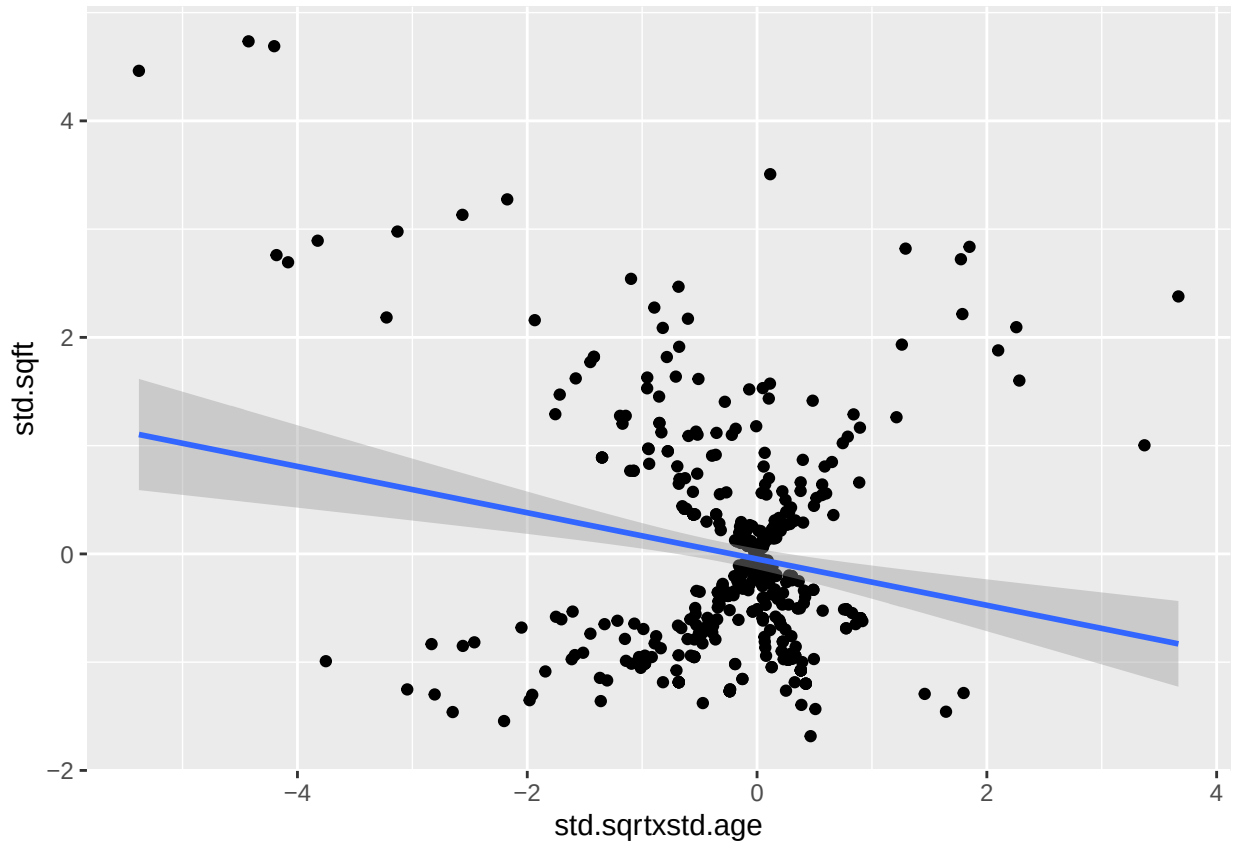
```
SLO %>% ggplot(aes(x=SqftxAge, y=Sqft.1000)) +
  geom_point() +
  geom_smooth(method='lm', se=TRUE)
```




(18) It turns out that one way to "control" this induced linear association with quantitative variables is to standardize the variables involved in the interaction first. Standardizing the variables does not change their association with the response variable, but it will change their association with each other. The standardized variables have also been included for you in the data file (std sqft, std age, and std sqft $\times$ std age).

Create a scatterplot of the interaction between the standardized variables and the standardized square footage variable.

```
SLO %>% ggplot(aes(x=std.sqrtxstd.age, y =std.sqft)) +
  geom_point()+
  geom_smooth(method='lm')
```




(19) Create a linear model with the standardized values including the interaction. How does the significance of the main effect change if we standardize the variables before including the interaction? (Include supporting evidence.)

```
std.lm <- SLO %>% lm(price~std.sqft * std.age, data=.)
summary(std.lm)
```

```
##
## Call:
## lm(formula = price ~ std.sqft * std.age, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7703 -0.7512 -0.1991  0.5868  7.7595
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.69494    0.07033  95.192  < 2e-16 ***
## std.sqft          2.29569    0.07061  32.512  < 2e-16 ***
## std.age           0.58283    0.07069   8.244 1.84e-15 ***
## std.sqft:std.age -0.28271    0.07461  -3.789 0.000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.447 on 450 degrees of freedom
## Multiple R-squared:  0.7249, Adjusted R-squared:  0.723
## F-statistic: 395.2 on 3 and 450 DF,  p-value: < 2.2e-16
```

```
SLO_std = SLO %>%
  mutate_at(list(~scale(.)), .var = vars(Sqft.1000, age))
std.lm2 <- SLO_std %>% lm(price~Sqft.1000 * age, data=.)
summary(std.lm2)
```

```
##
## Call:
## lm(formula = price ~ Sqft.1000 * age, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7703 -0.7512 -0.1991  0.5868  7.7595
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.70811    0.07035  95.349  < 2e-16 ***
## Sqft.1000      2.33716    0.07189  32.512  < 2e-16 ***
## age            0.58121    0.07075   8.215 2.28e-15 ***
## Sqft.1000:age -0.28782    0.07596  -3.789 0.000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.447 on 450 degrees of freedom
## Multiple R-squared:  0.7249, Adjusted R-squared:  0.723
## F-statistic: 395.2 on 3 and 450 DF,  p-value: < 2.2e-16
```

(20) Provide an interpretation in context of the slope coefficient of the standardized square footage variable in this two-variable model with an interaction term. (Hint: What happens in this prediction equation if we use year = 1977, the mean year of our dataset?)

**STEP 5: Formulate conclusions**

(21) Write a summary of your analysis, as if to a realtor, about the usefulness of your two-variable model in predicting initial selling prices in this market. You should include any limitations of your model and how it can be used. If the interaction is significant, include a discussion of the nature of the interaction in terms your realtor friend can understand.

**STEP 6: Look back and ahead.**

(22) Your realtor friend wants to include number of bedrooms in the model. Give one possible advantage and one possible disadvantage to doing so.

(23) Create a model that includes square feet, age, and number of bedrooms. Interpret the sign of the slope coefficient of this variable. Does it make sense in this context? Explain.

```
sfagebed.lm <- SLO %>% lm(price ~ std.sqft + std.age + Bedrooms, data = .)
summary(sfagebed.lm)
```

```
##
## Call:
## lm(formula = price ~ std.sqft + std.age + Bedrooms, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9722 -0.7996 -0.2142  0.6905  8.4782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.41087    0.35263  21.016   <2e-16 ***
## std.sqft     2.49157    0.10080  24.717   <2e-16 ***
## std.age      0.66541    0.07528   8.840   <2e-16 ***
## Bedrooms    -0.22261    0.11910  -1.869   0.0623 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.464 on 450 degrees of freedom
## Multiple R-squared:  0.7183, Adjusted R-squared:  0.7164
## F-statistic: 382.4 on 3 and 450 DF,  p-value: < 2.2e-16
```

```
anova(sfagebed.lm)
```

```
## Analysis of Variance Table
##
## Response: price
##            Df  Sum Sq Mean Sq   F value  Pr(>F)
## std.sqft    1 2288.48 2288.48 1067.8224 < 2e-16 ***
## std.age     1  162.89  162.89   76.0045 < 2e-16 ***
## Bedrooms    1    7.49    7.49    3.4934 0.06226 .
## Residuals 450  964.41    2.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```