Simple Linear Regression

Lecture 6 Handout Solutions

Statistics 139

Topics

- Linear Regression
- Correlation and R^2
- Outliers

The material in this lab corresponds to the Lecture 6 Notes.

In this lab we will explore recent real estate sales (last 6 months) in the greater Harvard Square area. We'll investigate what variables may be useful to predict the selling price of residential homes. More specifically:

1. How does selling price of homes relate to the size of the home (measured by floor space).
2. Are Real Estate Agents correct when they say "location, location, location!" matters when determining home values.
3. How does type of home relate to selling price and the relationships above?

The data set 'harvardsqhomes.csv' contains several variables measured on homes sold in and around Cambridge, MA for June 25 - September 20 this year. Data come from Redfin.com. Variables useful for today's handout include:

- `price`: the selling price of the home, in US $.
- `sqft`: the living area of the home as measured by floor space, in square feet
- `type`: a categorical variable indicating whether the house is a condo, townhouse, single-family home, or multi-family home.
- `latitude`: the latitude of the property location, measured in 'decimal' form in relationship to the equator (positive means northern hemisphere).
- `longitude`: the longitude of the property location, measured in 'decimal' form in relationship to the prime meridian (negative means west of the prime meridian).

Concept Checks

a) If the predictor and the response were switched in a simple regression model, would the new estimated slope just be the reciprocal of the original? Why or why not?

Based on the formula $\hat{\beta}_1 = r_{xy}\frac{s_x}{s_y}$, if the X and Y variables are flipped, correlation is unchanged, but the ratio of standard deviations does flip. Thus the slope is **not** simply the reciprocal (unless the estimated correlation is 1 or -1… a perfectly fit line). Conceptually: by flipping the response and predictor, the minimization perspective changes: instead of minimizing vertical distances, it's like we are minimizing horizontal distances.

b) When can R^2 be negative? Can an OLS (ordinary least squares) regression model have an R^2 less than zero? Why or why not?

R^2 can be negative for a model in general if it performs worse than the horizontal line at $\bar{y}$. An example: if the scatterplot displays a positive association but our model has a negative slope. This will never happen under OLS: the worst case setting is a slope of zero and an intercept at $\bar{y}$. Note: R^2 can also be negative in a left-out set of data (validation or testing sets), which is a common approach in prediction modeling (aka, machine learning).

c) The OLS estimate of variance is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}{n-2} = \frac{SSE}{df_E}$$

.

Why the $n-2$ (where does this come from)? What is the sampling distribution of $\hat{\sigma}^2$?

$\hat{\sigma}^2 \sim \left(\frac{\sigma^2}{n-2}\right)\chi^2_{n-2}$. $n-2$ can be justified from many perspectives. This leads to the estimate being unbiased for the true σ. This is the result because of the *degree of freedom* that the sum of squares in the numerator has: in order to calculate this sum of squares error (around the line), the slope and intercept has to first be calculated. This results in the vector of Y being *anchored* at these two estimates, *eating up* those 2 degree of freedom. More conceptually: a line fit to $n = 2$ is guaranteed to go through the 2 points exactly (unless they have the same value of x), and thus it is impossible to estimate variability around the line. This formula agrees with that intuition: $\hat{\sigma}^2$ is undefined unless $n \geq 3$.

Question 1: Exploratory Analysis: Price and Size

a) Explore the data set and investigate each of the variables through the `summary` command. Comment on what you see especially anything that appears unusual.

There is a very long right tail with many outliers: 17 beds, 9.5 baths, and over 8,700 square feet!

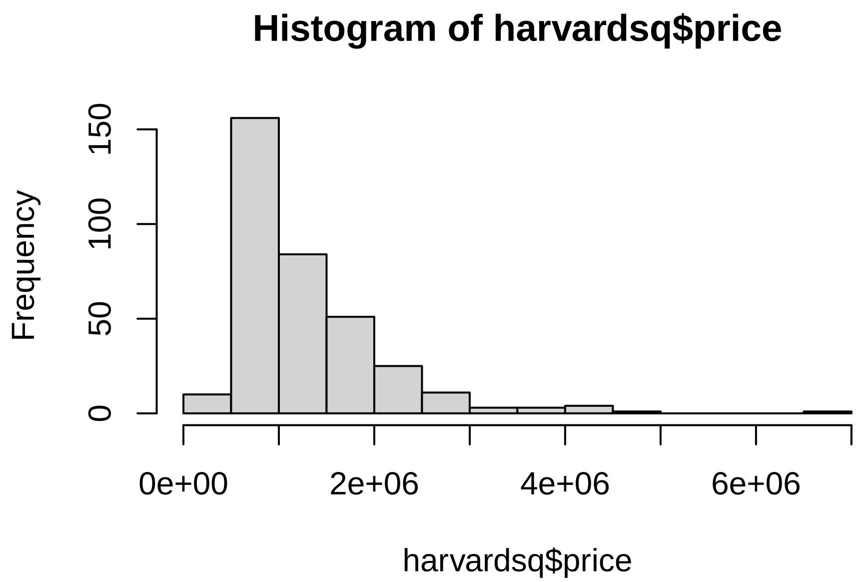
```
# look at some histograms
harvardsq = read.csv('data/harvardsqhomes.csv')

summary(harvardsq)
```

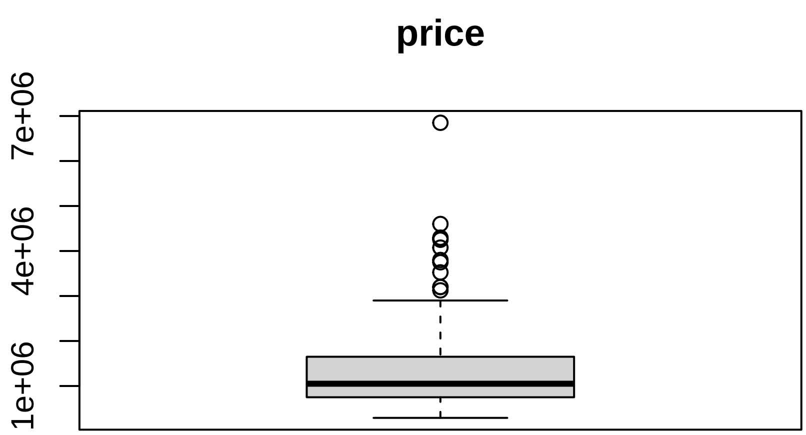
```
##      date               type              address              city          
##  Length:349         Length:349         Length:349         Length:349        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##
##
##
##
```

```
##       zip            price              beds            baths
##  Min.   :2138   Min.   : 291500   Min.   : 0.000   Min.   :1.000
##  1st Qu.:2139   1st Qu.: 750000   1st Qu.: 2.000   1st Qu.:1.000
##  Median :2139   Median :1050000   Median : 2.000   Median :2.000
##  Mean   :2141   Mean   :1291537   Mean   : 2.782   Mean   :2.107
##  3rd Qu.:2141   3rd Qu.:1650000   3rd Qu.: 3.000   3rd Qu.:2.500
##  Max.   :2414   Max.   :6850000   Max.   :17.000   Max.   :9.500
##
##  neighborhood.9          sqft         lotsize           year
##  Length:349         Min.   : 294   Min.   : 1025   Min.   :1805
##  Class :character   1st Qu.: 871   1st Qu.: 2092   1st Qu.:1894
##  Mode  :character   Median :1191   Median : 3049   Median :1915
##                     Mean   :1551   Mean   : 3608   Mean   :1931
##                     3rd Qu.:1897   3rd Qu.: 4042   3rd Qu.:1982
##                     Max.   :8737   Max.   :13873   Max.   :2022
##                                    NA's   :278
##       hoa              url                 mls               latitude
##  Min.   :     81   Length:349         Min.   :72809964   Min.   :42.36
##  1st Qu.:    231   Class :character   1st Qu.:72953839   1st Qu.:42.37
##  Median :    341   Mode  :character   Median :72969397   Median :42.37
##  Mean   : 18858                       Mean   :72970006   Mean   :42.37
##  3rd Qu.:    505                      3rd Qu.:72989817   3rd Qu.:42.37
##  Max.   :999999                       Max.   :73026427   Max.   :42.38
##  NA's   :132                          NA's   :60
##    longitude
##  Min.   :-71.13
##  1st Qu.:-71.11
##  Median :-71.10
##  Mean   :-71.10
##  3rd Qu.:-71.10
##  Max.   :-71.08
##
```

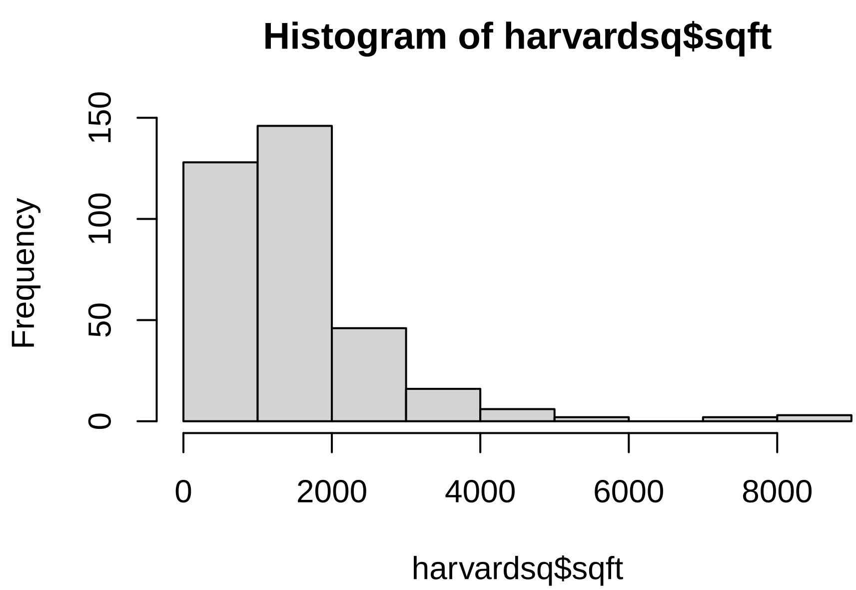
```r
hist(harvardsq$price)
```



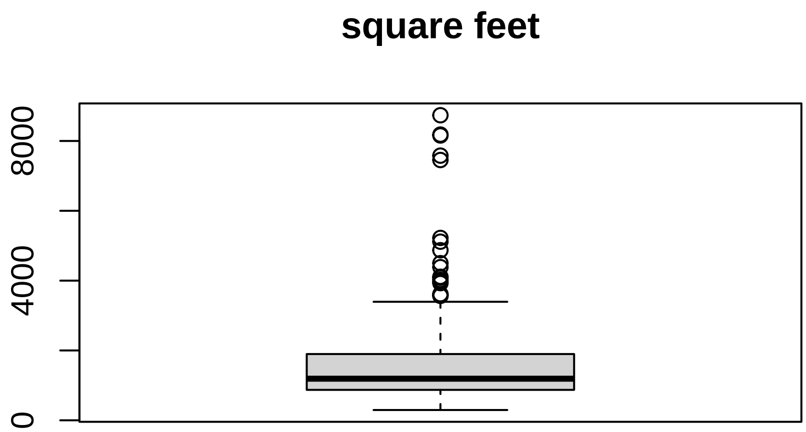

```
boxplot(harvardsq$price, main="price")
```



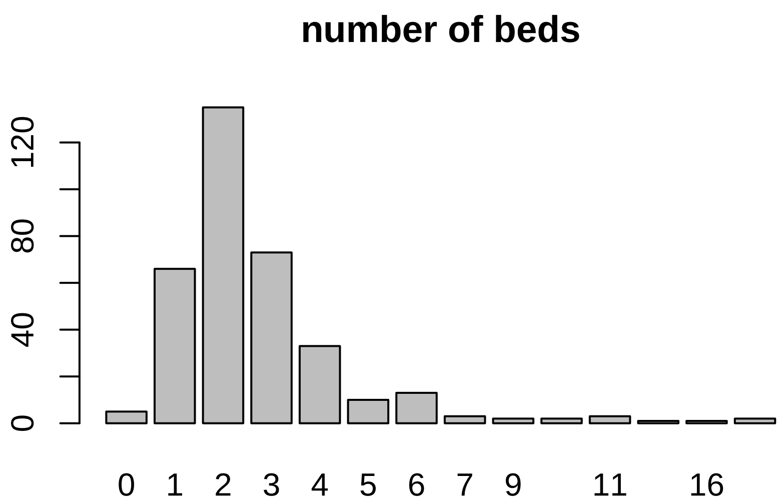

```
hist(harvardsq$sqft)
```



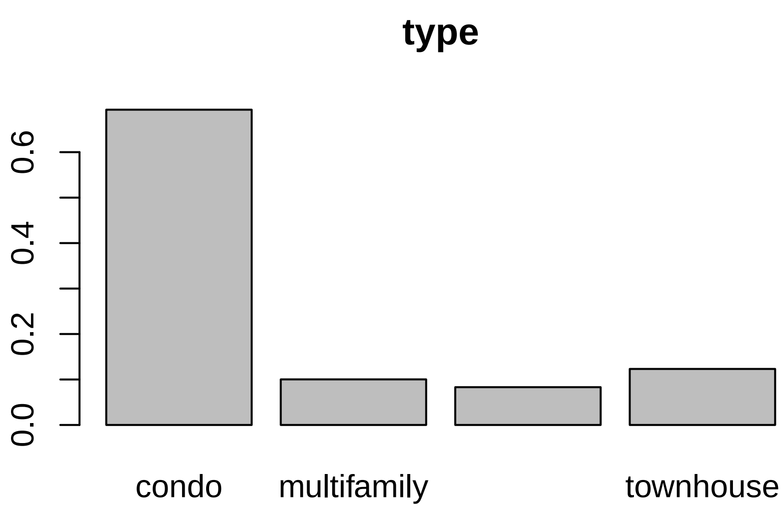
```
boxplot(harvardsq$sqft, main="square feet")
```



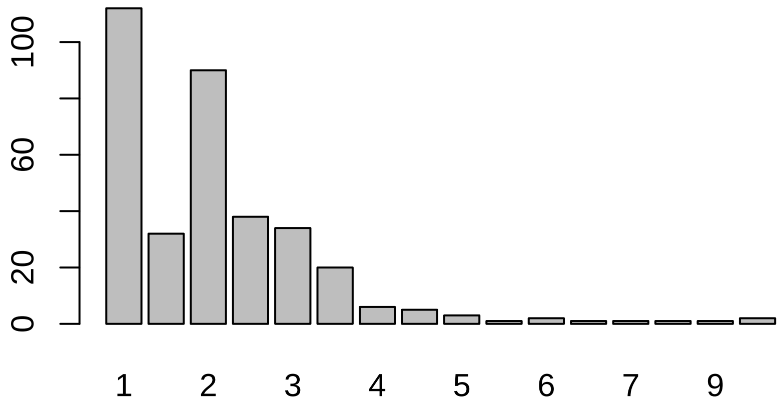

```
barplot(table(harvardsq$beds), main="number of beds")
```




```
barplot(prop.table(table(harvardsq$type)), main="type")
```




```
barplot(table(harvardsq$baths))
```




```r
#we have used summary() to look at individual variables,
#but you can give it an entire dataset as well
summary(harvardsq)
```

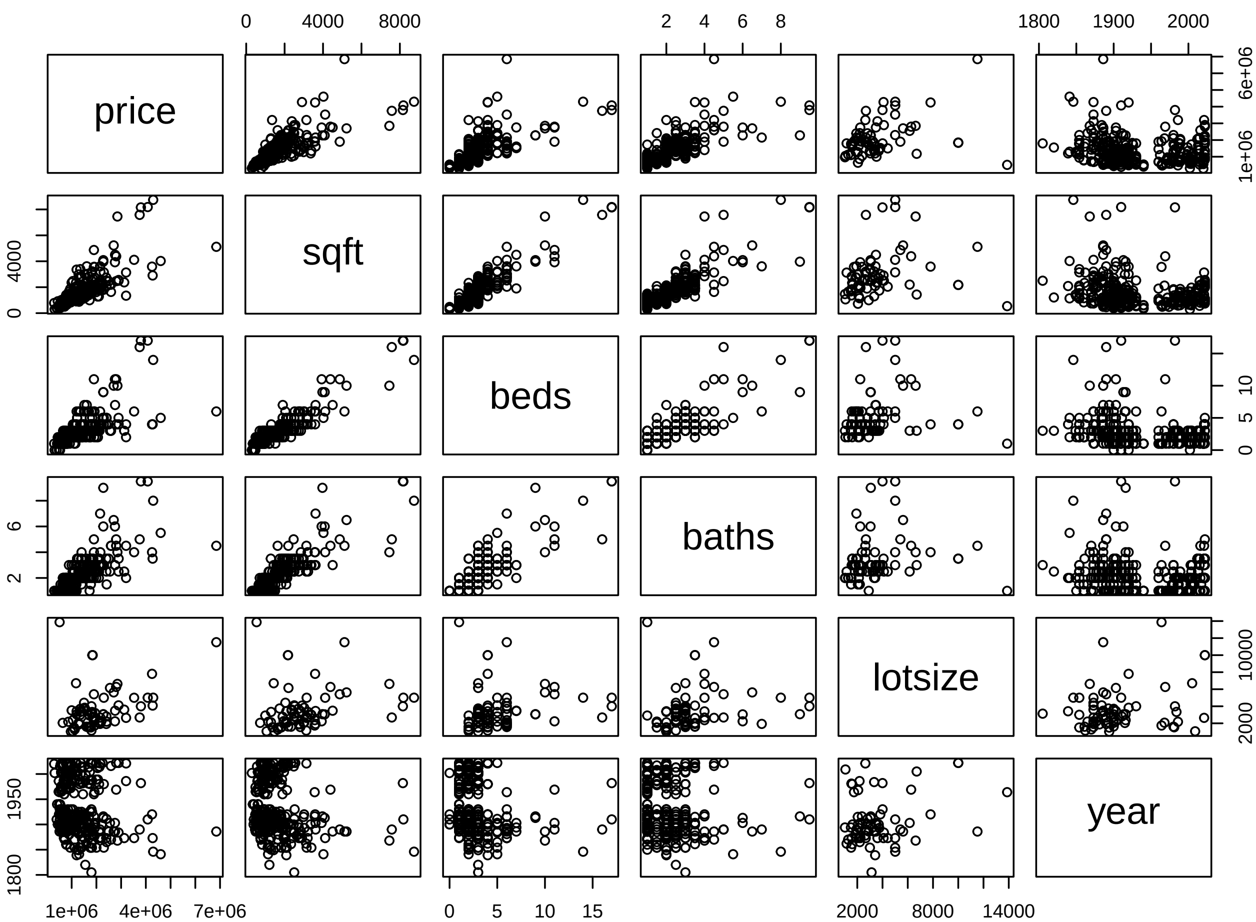
```
##      date               type             address              city          
##  Length:349         Length:349         Length:349         Length:349        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##       zip           price              beds            baths      
##  Min.   :2138   Min.   : 291500   Min.   : 0.000   Min.   :1.000  
##  1st Qu.:2139   1st Qu.: 750000   1st Qu.: 2.000   1st Qu.:1.000  
##  Median :2139   Median :1050000   Median : 2.000   Median :2.000  
##  Mean   :2141   Mean   :1291537   Mean   : 2.782   Mean   :2.107  
##  3rd Qu.:2141   3rd Qu.:1650000   3rd Qu.: 3.000   3rd Qu.:2.500  
##  Max.   :2414   Max.   :6850000   Max.   :17.000   Max.   :9.500  
##                                                                   
##  neighborhood.9          sqft         lotsize             year     
##  Length:349         Min.   : 294   Min.   : 1025   Min.   :1805  
##  Class :character   1st Qu.: 871   1st Qu.: 2092   1st Qu.:1894  
##  Mode  :character   Median :1191   Median : 3049   Median :1915  
##                     Mean   :1551   Mean   : 3608   Mean   :1931  
##                     3rd Qu.:1897   3rd Qu.: 4042   3rd Qu.:1982  
##                     Max.   :8737   Max.   :13873   Max.   :2022  
##                                    NA's   :278                   
##       hoa              url                 mls               latitude    
##  Min.   :    81   Length:349          Min.   :72809964   Min.   :42.36  
##  1st Qu.:   231   Class :character    1st Qu.:72953839   1st Qu.:42.37  
##  Median :   341   Mode  :character    Median :72969397   Median :42.37  
##  Mean   : 18858                       Mean   :72970006   Mean   :42.37  
##  3rd Qu.:   505                       3rd Qu.:72989817   3rd Qu.:42.37  
##  Max.   :999999                       Max.   :73026427   Max.   :42.38  
##  NA's   :132                          NA's   :60                        
```

```
##    longitude
##  Min.   :-71.13
##  1st Qu.:-71.11
##  Median :-71.10
##  Mean   :-71.10
##  3rd Qu.:-71.10
##  Max.   :-71.08
##
```

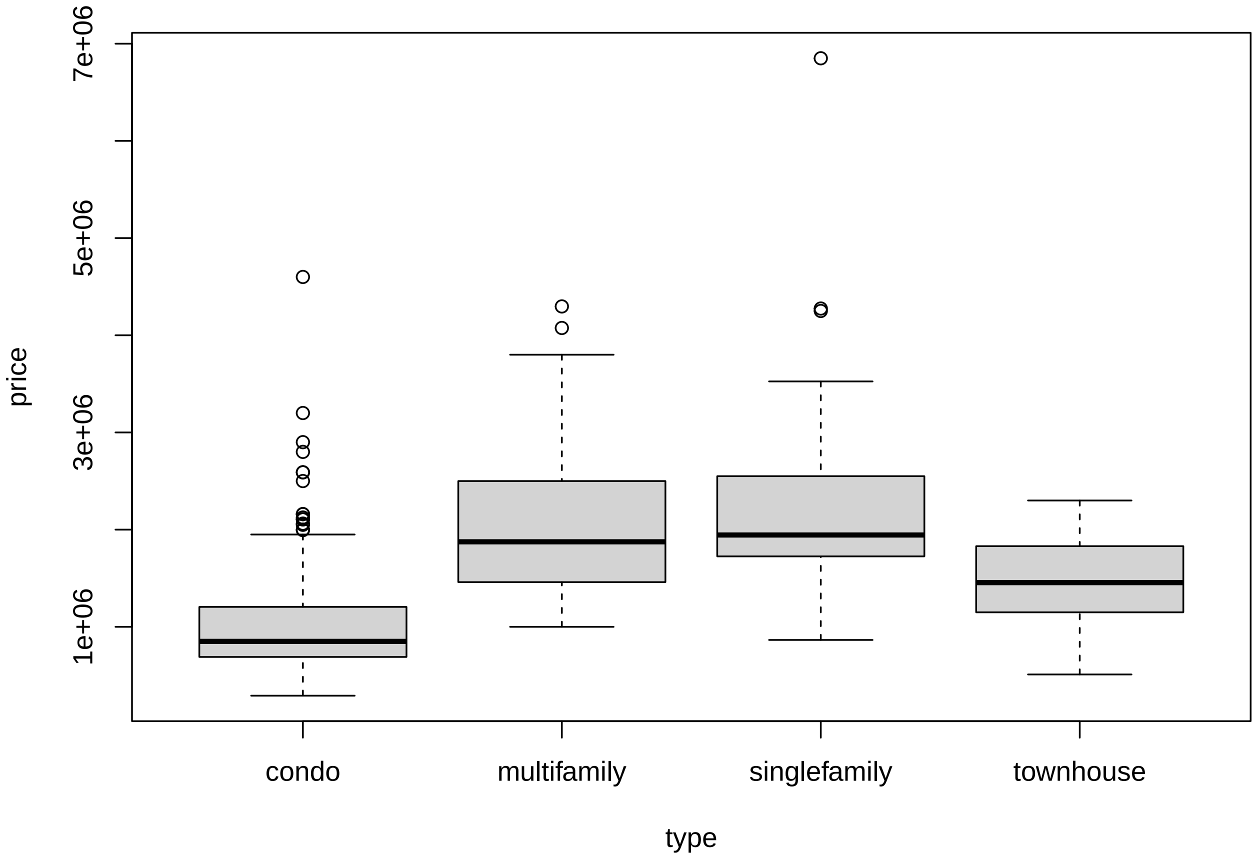
b) Visually investigate what variables may be related to selling price. What predictors may have the strongest association with `price`?

The variables sqft, beds, and baths, and lotsize have at least a moderate positive relationship with price. It's difficult to see a relationship with age (though there is an interesting two-cluster looking plot). Multi-family and single family homes are higher on average than condos and townhomes, and the neighborhoods are all over the place. . . though West Cambridge seems to be quite higher than the rest.

```
# look at some scatterplots
pairs(harvardsq[c('price','sqft','beds','baths','lotsize','year')])
```

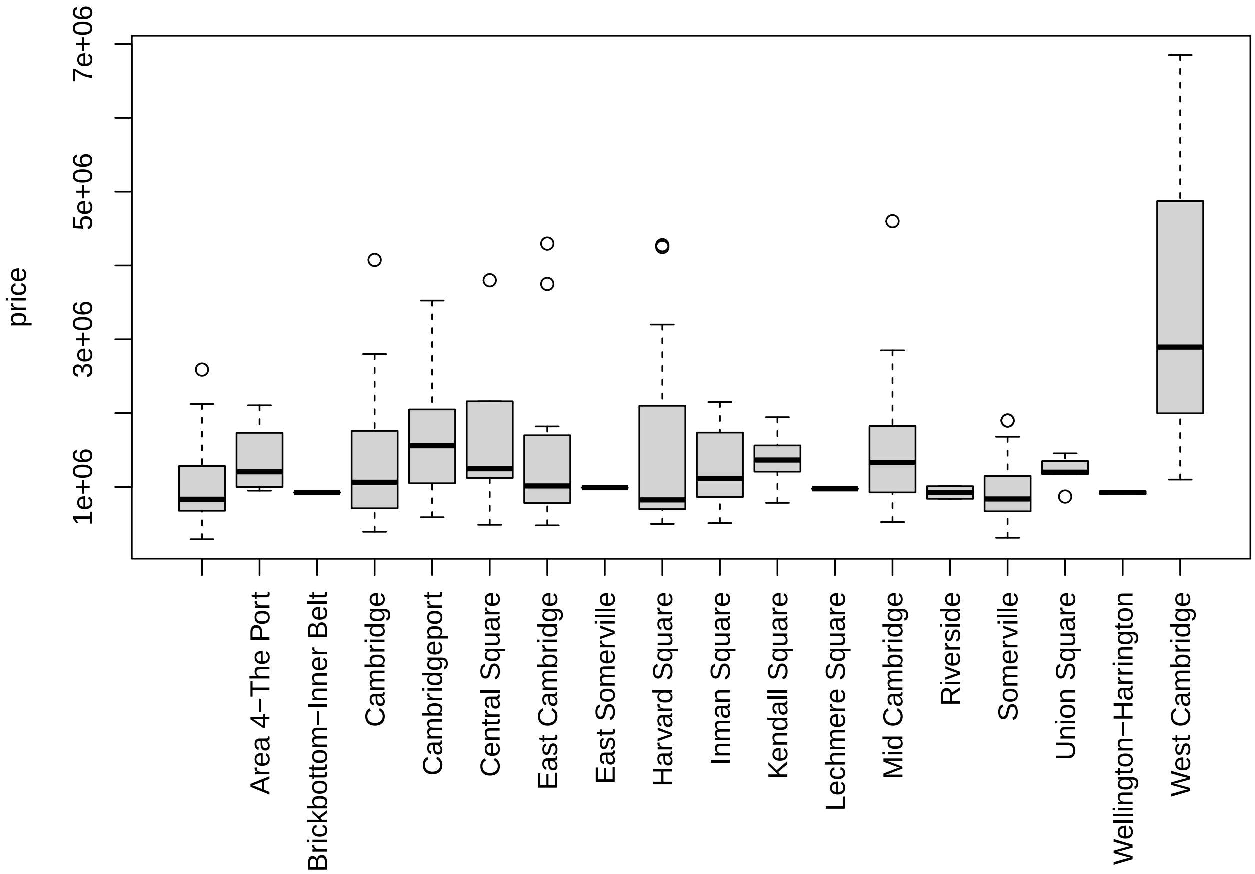


```r
boxplot(price~type,data=harvardsq)
```




```r
harvardsq$neighborhood = harvardsq$neighborhood.9

par(mar = c(10, 4, 4, 2) + 0.1)
boxplot(price~neighborhood,data=harvardsq, las=3, xlab="")
```



c) Fit the regression model to predict `price` from `sqtft` and plot the associated scatterplot. Add the estimated regression line to the plot. Interpret the estimates of $\beta_0, \beta_1, \sigma_2,$ and R^2.

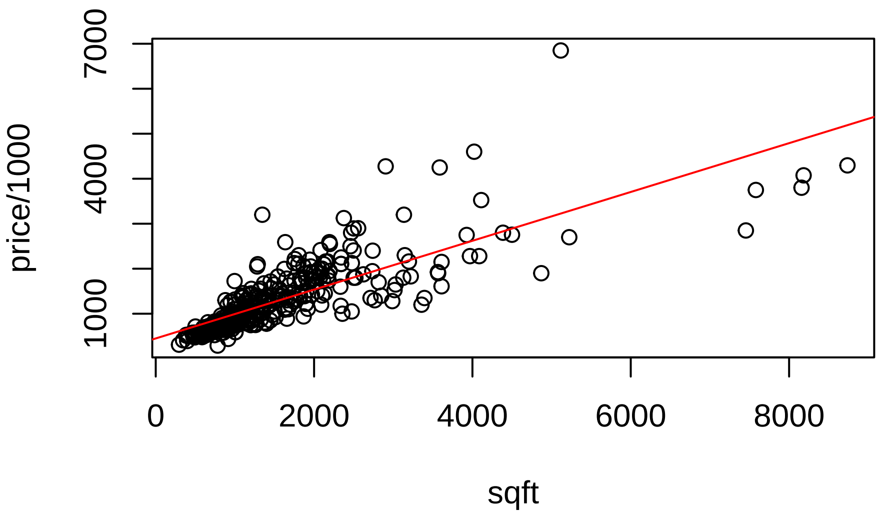
$\hat{\beta}_0$ is estimated to be 450.27852: the average selling price of properties without a house on them is roughly \$450,000 (do not read too much into this as it is an extrapolation). $\hat{\beta}_1$ is estimated to be 0.54253: an extra square foot of floor space is associated with a \$542.5 increase in selling price on average. $\hat{\sigma}^2$ is estimated to be 478.7: the standard deviation of individual home prices around their estimated average based on square footage is \$478,700. $R^2 = 0.6368$: 63.7% of the variability in selling prices of homes in the sample can be explained by square footage.

```r
# lm model and scatterplot with fitted line
summary(lm1 <- lm(price/1000~sqft,data=harvardsq))
```

```
##
## Call:
## lm(formula = price/1000 ~ sqft, data = harvardsq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1644.3  -231.9   -97.3   152.8  3624.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 450.27852   42.65951   10.55   <2e-16 ***
```

```
## sqft          0.54253    0.02199   24.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 478.7 on 347 degrees of freedom
## Multiple R-squared:  0.6368, Adjusted R-squared:  0.6358
## F-statistic: 608.4 on 1 and 347 DF,  p-value: < 2.2e-16
```

```
plot(price/1000~sqft,data=harvardsq)
abline(lm1,col="red")
```



d) There are 5 unusually large homes in the scatter plot in the previous part (`sqft` > 7000).

(i) What will happen to the four statistics listed above ($\beta_0, \beta_1, \sigma_2,$ and R^2) if these observations were and the regression model was refit?

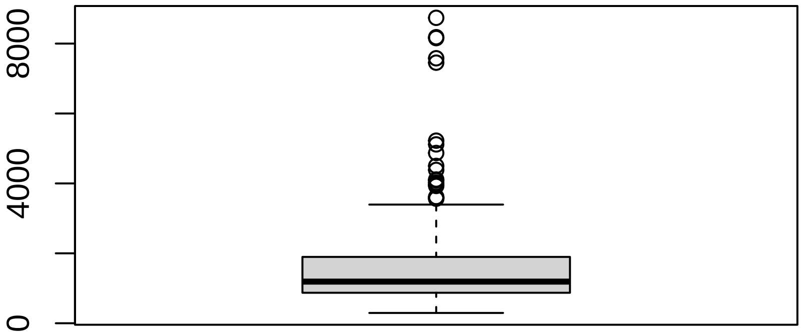
These observations are in the lower-righthand side of the plot. Since the points are all below the line, their removal will lead to an increase in the slope, leading to a decrease in the intercept. R^2 will likely go up because even though these observation are majors source of variability in 'price' (so the total sums of squares, SST, will be smaller after their removal: the denominator in the R^2 formula), these observation are not explained all that well (have relatively large magnitude in the residuals), thus these points are likely contributing a lot to the sums of squares error, SSE, in the numerator $R^2 = 1 - \frac{SSE}{SST}$. Since they have a fairly large distance from the line, the remaining residuals are likely to have lower spread on average.

(ii) Provide 1-2 sentences to justify when removing these observations might be reasonable.

Removing these observations would only be reasonable if, for example, it turned out these properties were not actually residential properties like the others. This would require, for example, investigating the listing closely, and/or getting expert opinion from domain experts (i.e., realtors in this setting). You should not remove an outlier simply because it is an outlier based on the distributions of your variables of interest, as that opens you up to the (valid) criticism that you are manipulating the data to come to a specific conclusion. You could, however, do a sensitivity analysis with the outliers removed to provide additional perspective.

In this case, after inspecting the listing, it seems these outliers are acutally apartment complexes, not standard residential homes like the others.

```r
boxplot(harvardsq$sqft)
```




```r
harvardsq[harvardsq$sqft > 6000,]
```

```
##                date        type            address      city  zip   price beds
## 251 August-9-2022 multifamily  347-349 Broadway Cambridge 2139 2850000   10
## 255  July-27-2022 multifamily 166-168 Auburn St Cambridge 2139 3800000   17
## 273 August-5-2022 multifamily   93 Thorndike St Cambridge 2141 4296675   14
## 275   July-8-2022 multifamily  337 Cambridge St Cambridge 2141 3750000   16
## 276   June-1-2022 multifamily  359 Cambridge St Cambridge 2141 4075000   17
##     baths neighborhood.9 sqft lotsize year hoa
## 251   4.0  Mid Cambridge 7454    6624 1868  NA
## 255   9.5 Central Square 8157    3998 1982  NA
## 273   8.0 East Cambridge 8737    5007 1846  NA
## 275   5.0 East Cambridge 7580    2685 1890  NA
## 276   9.5      Cambridge 8183    5006 1910  NA
##                                                                          url
## 251    https://www.redfin.com/MA/Cambridge/347-Broadway-02139/home/179444100
## 255    https://www.redfin.com/MA/Cambridge/166-Auburn-St-02139/home/11561865
## 273  https://www.redfin.com/MA/Cambridge/93-Thorndike-St-02141/home/11552523
## 275 https://www.redfin.com/MA/Cambridge/337-Cambridge-St-02141/home/11552225
## 276 https://www.redfin.com/MA/Cambridge/359-Cambridge-St-02141/home/11552079
##          mls latitude longitude   neighborhood
## 251 72966794 42.37032 -71.10326  Mid Cambridge
## 255 72935701 42.36350 -71.10526 Central Square
## 273 72908566 42.37013 -71.08232 East Cambridge
## 275 72893617 42.37140 -71.08163 East Cambridge
## 276 72951001 42.37151 -71.08202      Cambridge
```

e) Remove the unusual points from the scatter plot in the previous parts. Call this new data frame `harvardsq.clean` and refit the regression model on this smaller data set. Does this updated model support your thoughts for the effect on $(\beta_0, \beta_1, \sigma_2,$ and $R^2)$?

The results are shown below: while the changes in the slope, intercept, and variance estimates match our suspicions, the change in R^2 does not. These houses' inclusions were affecting SST (of price) more than SSE, proportionally.}

```r
#remove the outlier and save resulting data.frame as 'harvardsq.clean'
harvardsq.clean = harvardsq[harvardsq$sqft < 7000, ]

summary(lm2 <- lm(price/1000~sqft,data=harvardsq.clean))
```

```
##
## Call:
## lm(formula = price/1000 ~ sqft, data = harvardsq.clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1654.09  -173.73   -52.27   130.00  3130.28
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 275.01301   47.24113   5.821 1.34e-08 ***
## sqft          0.67332    0.02785  24.179  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449.2 on 342 degrees of freedom
## Multiple R-squared:  0.6309, Adjusted R-squared:  0.6298
## F-statistic: 584.6 on 1 and 342 DF,  p-value: < 2.2e-16
```

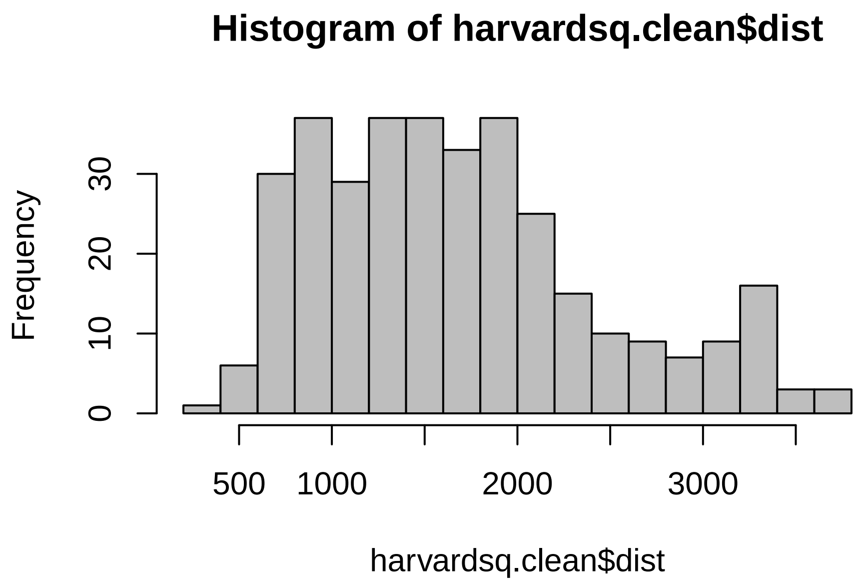
f) Do a little feature engineering. Use the function `distHaversine` within the `geosphere` package to create a variable called `dist` within `harvardsq.clean` which estimates the distance of each property from the Harvard Sq T stop.

The histogram is shown below: distance is measured in meters, the median distance from the Harvard Square T stop is 1.6 kilometers, ranging from a bit below 0.5 km up to a bit above 3.5 km.

```r
#install.packages(c("geosphere"))
library(geosphere)
```

```
## The legacy packages maptools, rgdal, and rgeos, underpinning the sp package,
## which was just loaded, will retire in October 2023.
## Please refer to R-spatial evolution reports for details, especially
## https://r-spatial.org/r/2023/05/15/evolution4.html.
## It may be desirable to make the sf package available;
## package maintainers should consider adding sf to Suggests:.
## The sp package is now running under evolution status 2
##      (status 2 uses the sf package in place of rgdal)
```

```r
harvardsq.loc = c(-71.1190,42.3736)
#this is the location (longitude, latitude) of Harvard
#Square T stop
harvardsq.clean$dist = distHaversine(harvardsq.clean[c('longitude','latitude')],
                                     harvardsq.loc)
hist(harvardsq.clean$dist,col="gray",breaks=20)
```




```r
summary(harvardsq.clean$dist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   359.3  1068.4  1591.6  1678.7  2039.0  3608.4
```

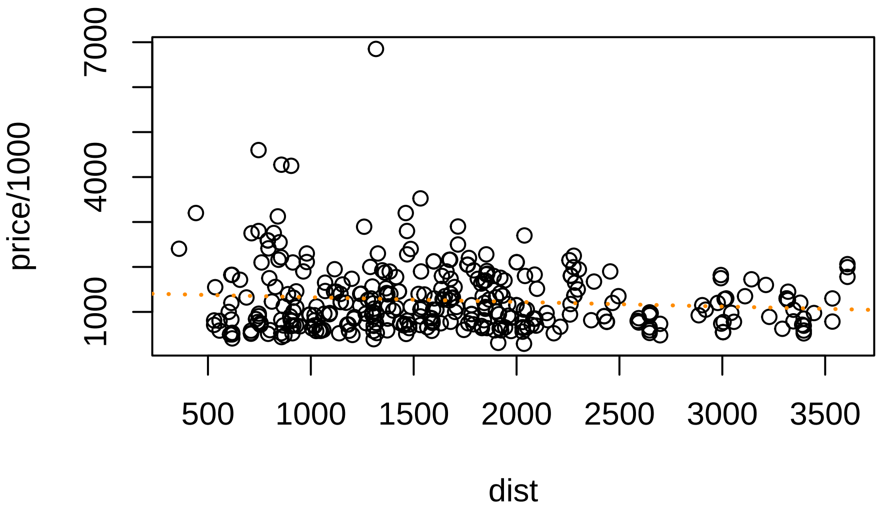
g) Fit a regression model to predict `price` from `dist` using `harvardsq.clean`. Interpret the results and provide a useful visual of the data and the model to illustrate these results.

Every meter further away from the Harvard Square T stop is associated with a \$ 103 decrease in the average selling price of a home. This is borderline significant ($t = -1.963,\ p = 0.051$), and it is a small effect (as it only explains 1.11% of the variability in prices).

```r
summary(lm3 <- lm(price/1000~dist,data=harvardsq.clean))
```

```
##
## Call:
## lm(formula = price/1000 ~ dist, data = harvardsq.clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -954.7 -525.8 -204.1  337.3 5557.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1428.16818   96.38715  14.817   <2e-16 ***
## dist          -0.10272    0.05234  -1.963   0.0505 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 735.3 on 342 degrees of freedom
## Multiple R-squared:  0.01114,    Adjusted R-squared:  0.008246
## F-statistic: 3.852 on 1 and 342 DF,  p-value: 0.0505
```

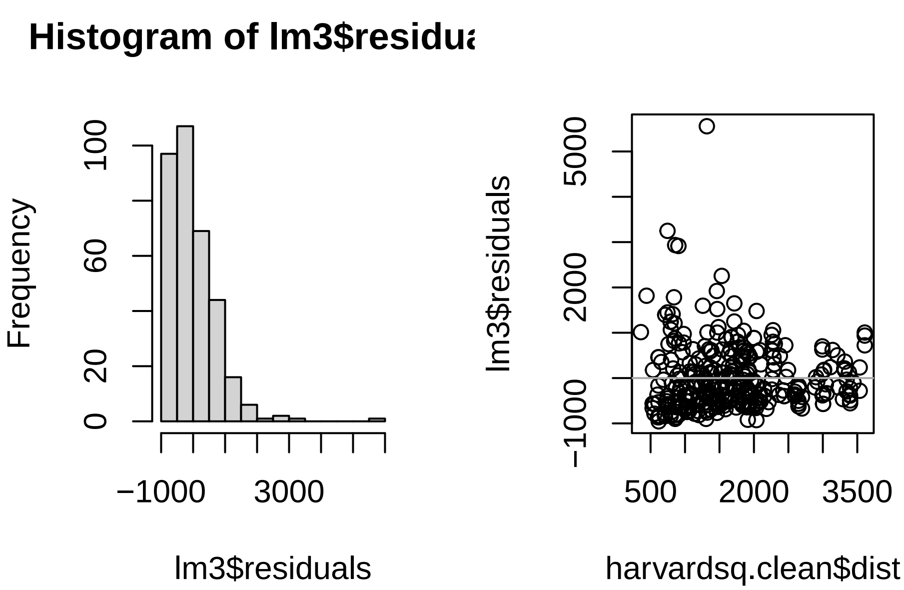
```r
plot(price/1000~dist,data=harvardsq.clean)
abline(lm3,col="darkorange",lty="dotted",lwd=2)
```



h) Comment on the appropriateness of the linear regression model in the two models run so far.

The relationship may not be linear in the previous one: it looks like it starts steeper and flattens off as distance increases. Plus the observations do not look normally distributed nor have constant variance around the estimated regression line. The residual plots below help make this determination.}

```r
par(mfrow=c(1,2))
hist(lm3$residuals)
plot(lm3$residuals~harvardsq.clean$dist)
abline(h=0, col="darkgray")
```

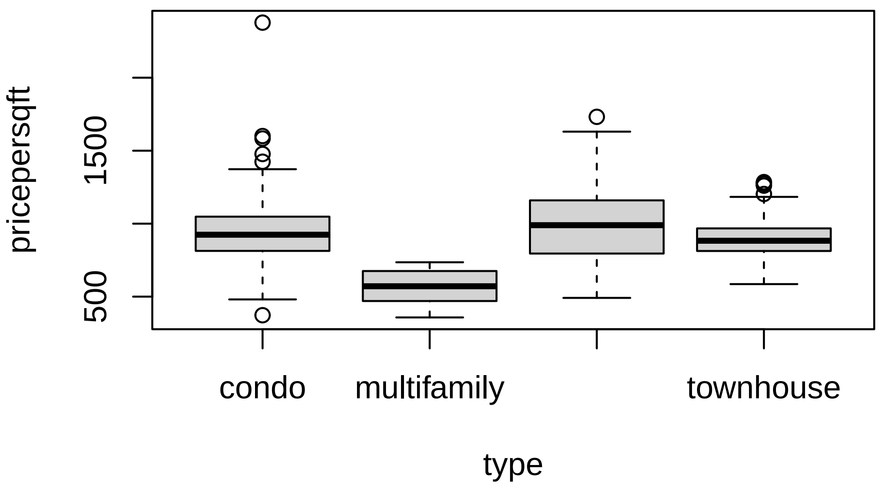


Question 2: How does type of home play a role?

a) Create a new variable `pricepersqft` in `harvardsq.clean` that measures the price per square foot for each property. Formally test whether selling price per square foot of floor space is significantly different across 'type' of property, and formerly determine which groups are different from one another. Provide a visual to support your conclusions.

Condos and single family homes have slightly higher cost per square foot on average than the other 3 types of properties (with condos having by far the highest outlier), with multi-family properties being the lowest. There is significant evidence that there is a difference between these groups on average (ANOVA $F = 29.38$, $p < 0.0001$). The pairwise t-tests suggest that multi-family homes are significantly less than all other types of homes, but all other comparisons are very similar after controlling for multiple comparisons via Bonferroni.}

```
harvardsq.clean$pricepersqft = harvardsq.clean$price/harvardsq.clean$sqft
boxplot(pricepersqft~type,data=harvardsq.clean)
```




```
summary(aov(pricepersqft~type,data=harvardsq.clean))
```

```
##              Df   Sum Sq Mean Sq F value Pr(>F)
## type          3  4059682 1353227   29.38 <2e-16 ***
## Residuals   340 15662939   46067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(harvardsq.clean$pricepersqft,harvardsq.clean$type,
                p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  harvardsq.clean$pricepersqft and harvardsq.clean$type
##
##              condo   multifamily singlefamily
## multifamily  < 2e-16 -           -
## singlefamily 0.34    6.8e-14     -
## townhouse    1.00    3.4e-10     0.24
##
## P value adjustment method: bonferroni
```

b) Fit two separate regression models: one for (condos and townhouses combined) and a separate one for (single-family homes and multi-family homes combined) to predict `price` from `sqft` using the `harvardsq.clean` data frame. Interpret the results.

The two model results are shown below: a 1 square foot increase in floor space is associated with a \$911.55 increase for condos and townhomes, while it is associated with just a \$490.3 increase in price for multi- and single-family homes. It would be interesting to see if this roughly \$420 difference in slope is statistically significant.}

```
# fit 2 separate regression models.  The argument 'subset'
# within the 'lm' command could be useful.

summary(lm4 <- lm(price/1000~sqft,data=harvardsq.clean,
                  subset=(type == 'condo' | type == 'townhouse')))
```

```
##
## Call:
## lm(formula = price/1000 ~ sqft, data = harvardsq.clean, subset = (type ==
##     "condo" | type == "townhouse"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -965.98 -141.85   -7.03   94.79 1961.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.59242   41.07669   0.282    0.778
## sqft         0.91155    0.03186  28.610   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279.6 on 283 degrees of freedom
## Multiple R-squared:  0.7431, Adjusted R-squared:  0.7422
## F-statistic: 818.6 on 1 and 283 DF,  p-value: < 2.2e-16
```

```
summary(lm5 <- lm(price/1000~sqft,data=harvardsq.clean,
                  subset=(type == 'singlefamily' | type == 'multifamily')))
```

```
##
## Call:
## lm(formula = price/1000 ~ sqft, data = harvardsq.clean, subset = (type ==
##     "singlefamily" | type == "multifamily"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1188.7  -524.6  -137.6   329.4  3640.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 701.1646   331.9211   2.112    0.039 *
## sqft          0.4903     0.1122   4.369 5.33e-05 ***
```

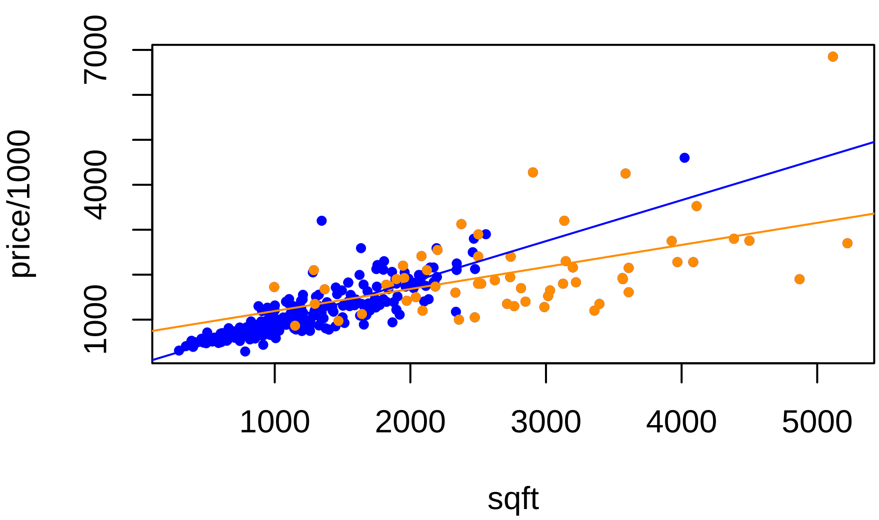
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 839 on 57 degrees of freedom
## Multiple R-squared:  0.2509, Adjusted R-squared:  0.2378
## F-statistic: 19.09 on 1 and 57 DF,  p-value: 5.333e-05
```

c) Create a well-chosen visual to illustrate the results of the previous model. Be sure to color-code the points and the lines to represent the 2 different groups. Interpret what you see.

The plot below depicts the (condos and townhomes) group in blue and the (multi- and single-family homes) in orange. Note the slope is higher in the 'condo' group, but they are actually predicted to sell for less for smaller homes (below 2000 square feet, where most of the 'condo' group is). They cross at about 1800 square feet.}

```r
# Create a single visual to illustrate these two models.
# The commands 'plot', 'points' and 'lines' could be useful.

plot(price/1000~sqft,data=harvardsq.clean,col="blue",cex=0.7,pch=16)
points(price/1000~sqft,data=harvardsq.clean,col="darkorange",
       subset = (type == 'singlefamily' | type == 'multifamily'),cex=0.7,pch=16)
abline(lm4,col="blue")
abline(lm5,col="darkorange")
```



d) How could you formally test if the slopes in the two subset models in part (b) are significantly different from one another? Note: we'll learn how to do this via two different approaches in future lectures (both mathematically equivalent).

There are at least two ways to do this: (1) the data could be combined into a single model with 3 predictors: X_1 = sqft, X_2 = a binary predictor to depict the two groups and X_3 = the interaction between the two and the t-test for the interaction term could be used to make this determination (we will learn how to do this in a couple weeks), or (2) a t-test for the difference in slopes from these two models, performed below (note, the two estimates are independent, so the variances simply add):

$$H_0 : \beta_{1,condos} - \beta_{1,single/multi} = 0$$

$$T = \frac{\hat{\beta}_{1,condos} - \hat{\beta}_{1,single/multi}}{\sqrt{\widehat{Var}(\hat{\beta}_{1,condos}) + \widehat{Var}(\hat{\beta}_{1,single/multi})}} = \frac{0.91155 - 0.4903}{\sqrt{0.03186^2 + 0.1122^2}} = 3.612$$

This t-statistic ($t = 3.612$) will be significant at any degrees of freedom since it is well above 1.96 (for a 2-sided test). The actual degrees of freedom (df) is not exact, but conservatively it should be the minimum of the two df from the models (283 and 57). The p-value is estimated to be 0.0006, and thus the result is significant (the slopes may truly be different in the two groups).}

```
print(t.stat <- (lm4$coef[['sqft']]-lm5$coef[['sqft']])/
           sqrt(summary(lm4)$coef[['sqft','Std. Error']]^2+
                  summary(lm5)$coef[['sqft','Std. Error']]^2))
```

```
## [1] 3.611979
```

```
2*(1-pt(abs(t.stat),df=min(summary(lm4)$df[2],summary(lm5)$df[2])))
```

```
## [1] 0.0006429778
```