Michigan Home Prices

Step 1: Ask a research question

1. I want to predict how much should I expect to pay for a home. What explanatory variables might explain variation in home prices?

SOLUTION: explanatory variables might include: location, square footage, number of baths/beds

Step 2: Design a study and collect data

2. ID the observational units and the response variable of interest. Is the response variable quantitative or categorical? What kinds of graphs and numerical summaries can you use to explore these data?

```
library(tidyverse)
home.data<-read.table("http://www.isi-stats.com/isi2/data/homeprices.txt",header=T, fill=T)
head(home.data)
```

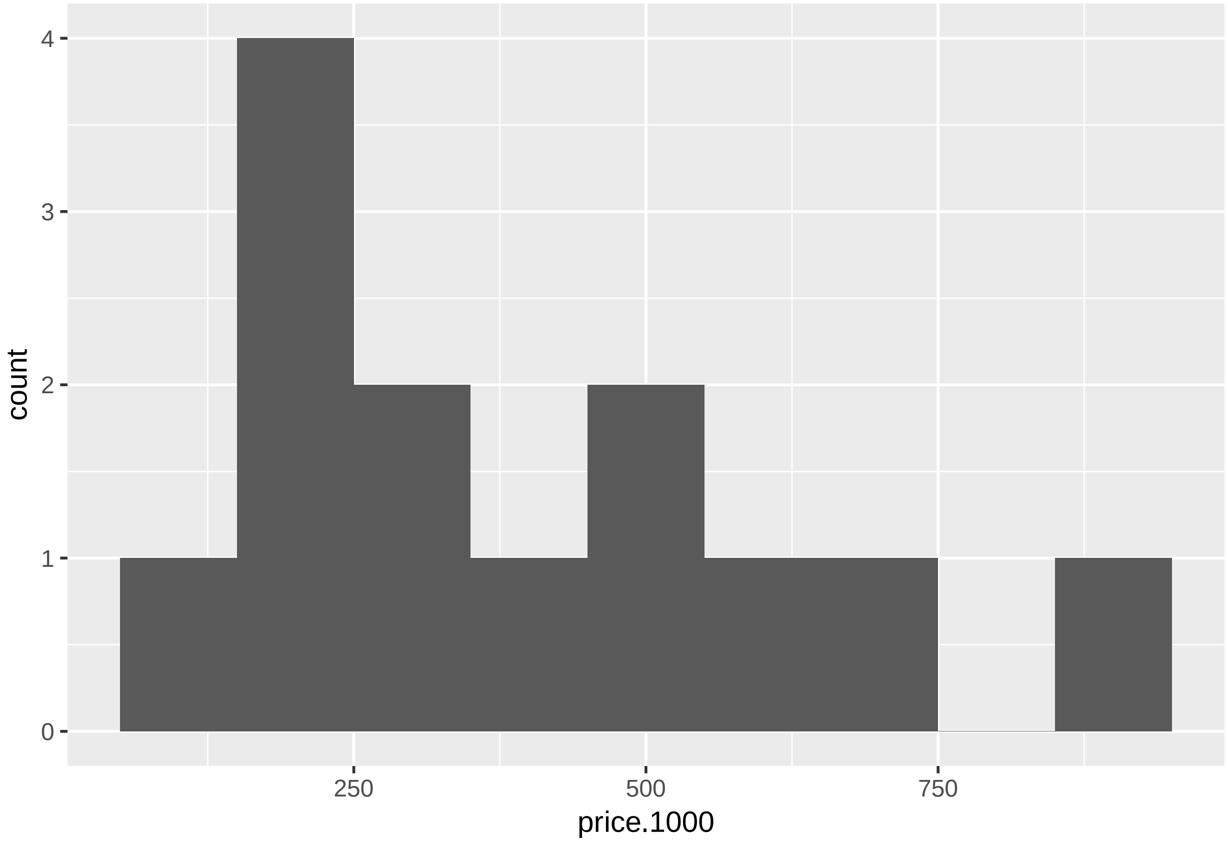
```
##   sqft price.1000      lake
## 1 2700      639.9 lakefront
## 2 3353      898.0 lakefront
## 3 1600      410.0 lakefront
## 4 1740      529.9 lakefront
## 5 1907      495.0 lakefront
## 6 3262      749.9 lakefront
```

SOLUTION: Observational units: homes north of Lake Macatawa in Michigan Response variable of interest: sale price Response variable is quantitative Can use a bar graph, histogram, mosaic graph, and scatter plot; also use descriptive statistics

Step 3: Explore the data

3. Describe the shape of the home prices in this sample

```
home.data %>%
  ggplot(aes(x=price.1000)) +
    geom_histogram(binwidth = 100)
```



```
summary(home.data$price.1000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   119.0   210.0   324.9   408.1   529.9   898.0
```

SOLUTION: The distribution of home prices in this samples is skewed to the right (more values below the mean than above the mean)

4. If we want to estimate a typical home price, would the mean be reasonable to use?

SOLUTION: No, because the data is skewed right, the mean would over-predict a typical home price. The median would be more appropriate (mean should only be used when the distribution is symmetric)

5. Do you think the mean is larger or smaller than the median.

SOLUTION: Clearly larger (skewed right and from summary statistics)

6. What is the standard deviation of the home prices? What are the units? Interpret this value. What are possible explanations for this variation?

```
sd(home.data$price.1000)
```

```
## [1] 240.9228
```

SOLUTION: Standard deviation is 240.9228; in other words, the typical distance of a home is $240,920 away from the mean. This will not provide an accurate predication of home prices. Possible explanations are size and location.

7. The first home in the data set sold for $639,000 - if we had used the mean to predict that price, how far off would we have been?

```
639-mean(home.data$price.1000)
```

```
## [1] 230.9385
```

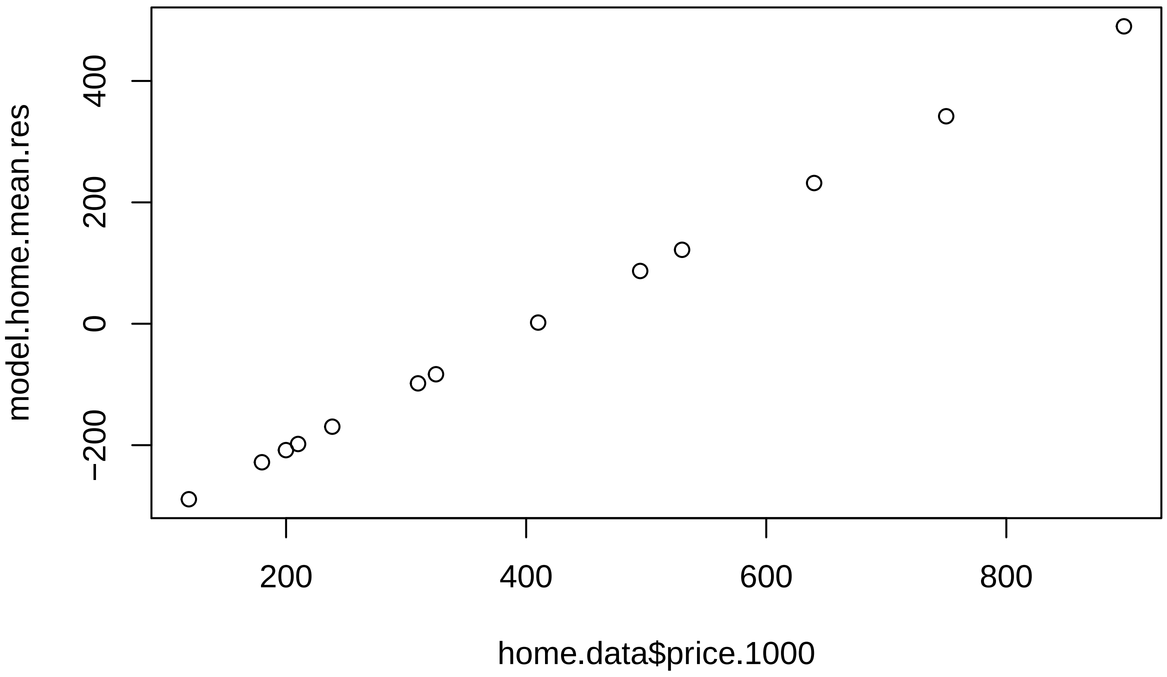
We would have been $230,938.50 below the actual price

8. Using the mean as our prediction, did we over-predict or under-predict the prices of the first house? (Is the residual positive or negative?)

SOLUTION: We under predicted, so the residual is positive (actual - predicted)

9. How many residuals are positive and how many are negative?

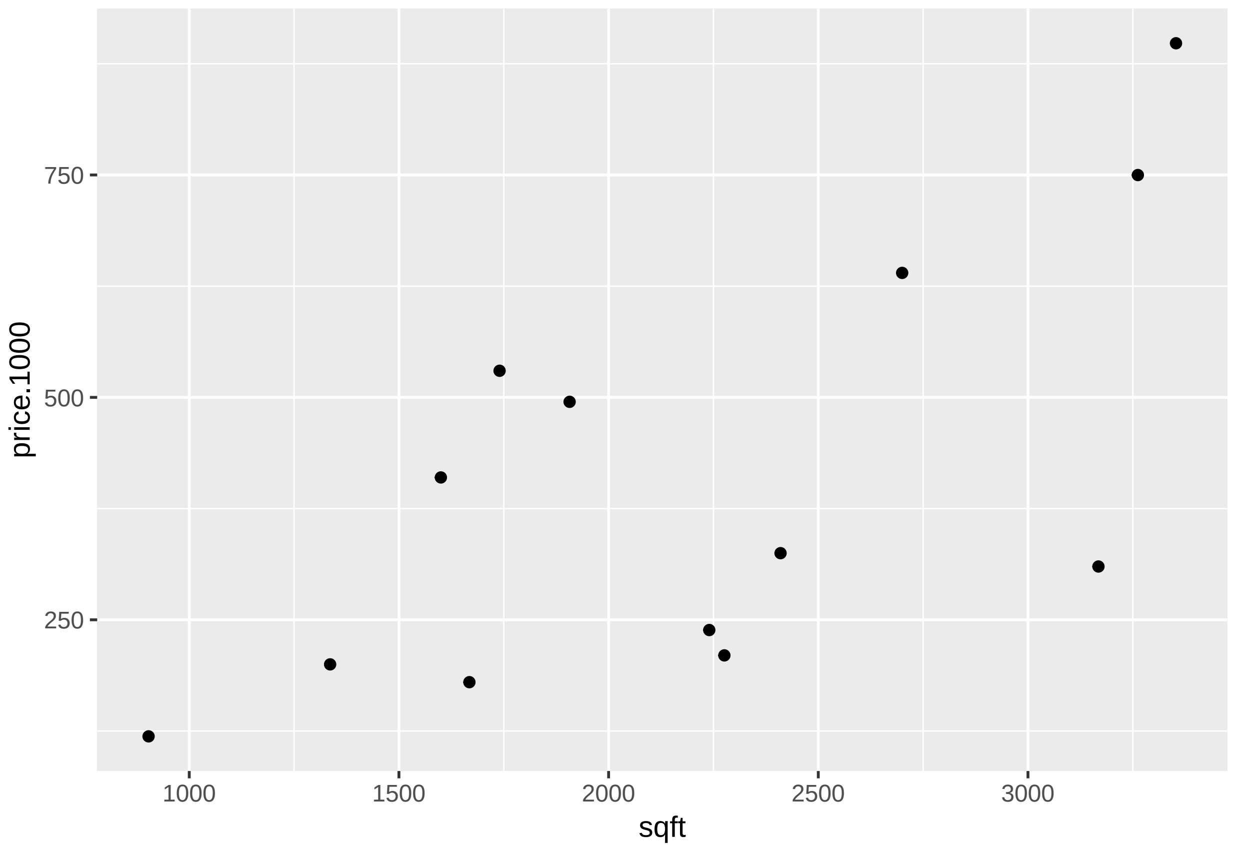
```
model.home.mean = lm(price.1000 ~ 1, data = home.data)
model.home.mean.res = resid(model.home.mean)
plot(home.data$price.1000, model.home.mean.res)
```



6 are positive (over-predicted) and 7 are negative (under-predicted)

10. Describe the association between home price and square footage (strength, direction, linearity). Does the association behave as you would have predicted? Explain

```
home.data %>%
  ggplot(aes(x=sqft, y=price.1000)) +
  geom_point()
```



```
model.home.1 = lm(price.1000 ~ sqft, data = home.data)
summary(model.home.1)
```

```
##
## Call:
## lm(formula = price.1000 ~ sqft, data = home.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304.70 -128.44  -13.74  128.98  244.04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.36870  161.36807  -0.368   0.7199
## sqft          0.21274    0.06963   3.055   0.0109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 185.1 on 11 degrees of freedom
## Multiple R-squared:  0.4591, Adjusted R-squared:  0.4099
## F-statistic: 9.335 on 1 and 11 DF,  p-value: 0.01094
```

SOLTUION: There is a moderate positive linear association between home price and square footage. A one-thousand increase in square footage is associated with a price increase of about $200,000

11a. What is the statistical model?

SOLUTION: Note here we will diverge slightly from our text and consider this the statistical model:

$$\begin{aligned} i &= \text{home} \\ y_i &= \text{Price, in thousands, of home } i \\ x_i &= \text{Square footage of home } i \\ y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\sim F(0, \sigma) \end{aligned}$$

The fitted model is:

Predicted price = -59.3687 + 0.21274*sqft; Standard error of residual = 185.1 thousand dollars

Note we use the generic F in our distribution of ϵ to denote that we are not making any normality assumptions at this point in time. 11b. Provide interpretations, in context, of the slope and intercept of the line. The intercept gives us the price of the home if the sqft were 0 (which is clearly not possible). The slope of the line gives us the $1,000 increase with each added square foot.

11c. What residual does this give you for the first home in the dataset? Is this residual smaller or larger than before?

```r
home.data$price.1000[1] - predict(model.home.1, data.frame(sqft = c(home.data$sqft[1])))
```

```
##        1 
## 124.8612
```

SOLUTION: The residual is 124.8612 ($124,861), which is less than before (using the mean, the residual was $230,940)

11d. Is the standard error of the residuals larger or smaller than before? Explain. summary(model.home.mean)

SOLUTION: the standard error of the residuals for the mean housing price was 240.9 thousand dollars and the standard error of the residuals for the linear model was 185.1 thousand dollars. Thus, it was larger before, indicating that the model accounting for square footage is a better model to predict home prices.

12. How do the two lines compare when considering location? Does it make sense?

```r
model.home.2 = lm(price.1000 ~ sqft*lake, data = home.data)
summary(model.home.2)
```

```
## 
## Call:
## lm(formula = price.1000 ~ sqft * lake, data = home.data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -54.16 -28.60 -14.15  29.64  73.93 
## 
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            86.76438   71.60612   1.212  0.25648    
## sqft                    0.21990    0.02831   7.769  2.8e-05 ***
## lakenotlakefront      -28.65098   91.32560  -0.314  0.76088    
## sqft:lakenotlakefront  -0.13595    0.03895  -3.491  0.00682 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 49.48 on 9 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9578
## F-statistic: 91.84 on 3 and 9 DF,  p-value: 4.547e-07
```

SOLUTION: The new statistical model is:

$$\begin{aligned} i &= \text{home}\\ y_i &= \text{Price, in thousands, of home}\\ x_{1,i} &= \text{Square footage of home } i,j\\ x_{2,i} &= 1 \text{ if house i is lakefront 0 otherwise}\\ y_{i,j} &= \beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\beta_3 x_{1,i}x_{2,i}\epsilon_i\\ \epsilon_i &\sim F(0,\sigma) \end{aligned}$$

From this model we get the two lines:

If Lake Front, 86.76 + 0.22(sqft) If NOT Lake Front - 58.11 + 0.084(sqft)

SE of Residuals for the model is: 49.48 thousand dollars

This model makes sense, because we would expect non-lakefront homes to start at a lesser value, and that as squarefoot increases the price would not increase as much as lakefront properties

13. New statistical model (See above).

SOLUTION: New Fitted Model: Predicted price = 86.764 + 0.2199(sqft) - 28.651(notLakeFront) - 0.136(sqft)(notLakeFront)

SE of Residuals: 49.48 thousand dollars

14. Use this model to predict the price of the first home.

```
predict(model.home.2, data.frame(sqft = c(home.data$sqft[1]), lake = as.factor(home.data$lake[1])))
```

```
##        1
## 680.4814
```

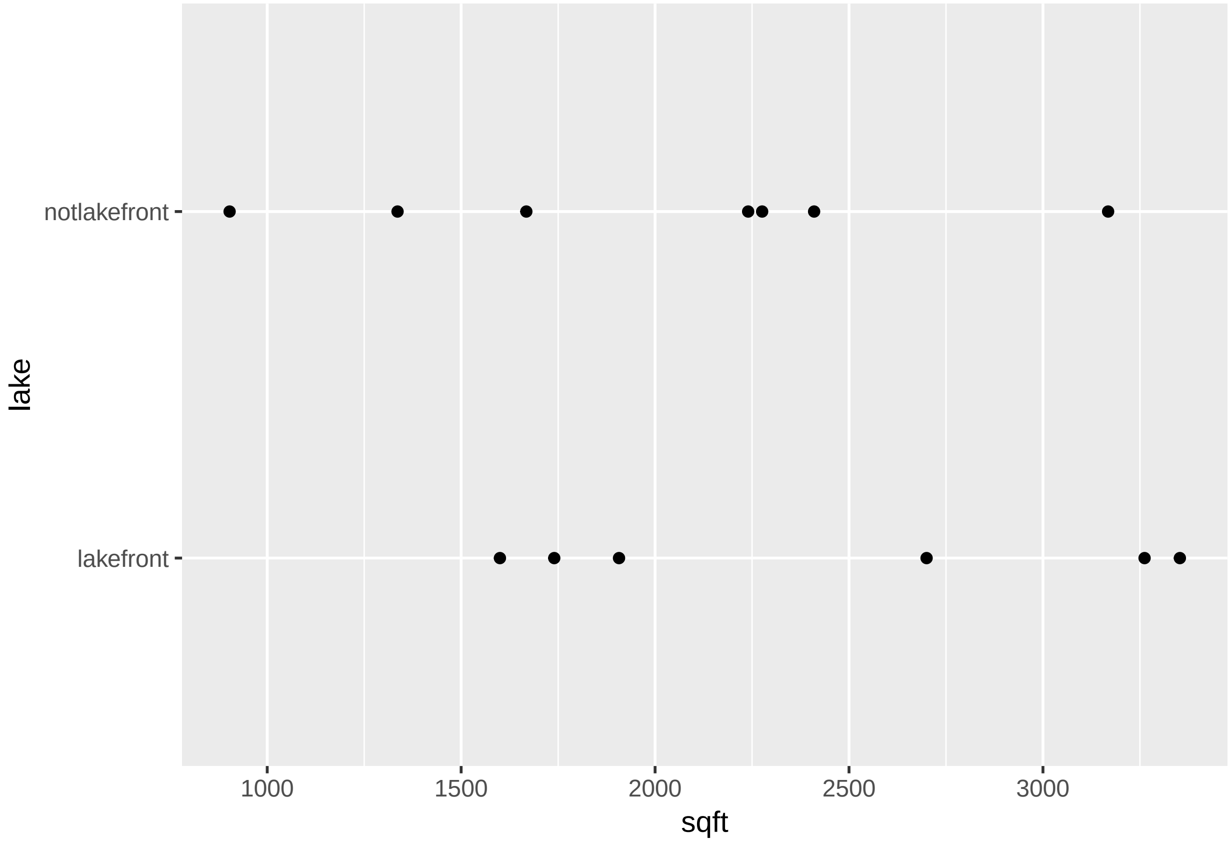
SOLUTION: The predicted home price is $680,481

15. Does including this additional location variable further reduce the SE of the residuals?

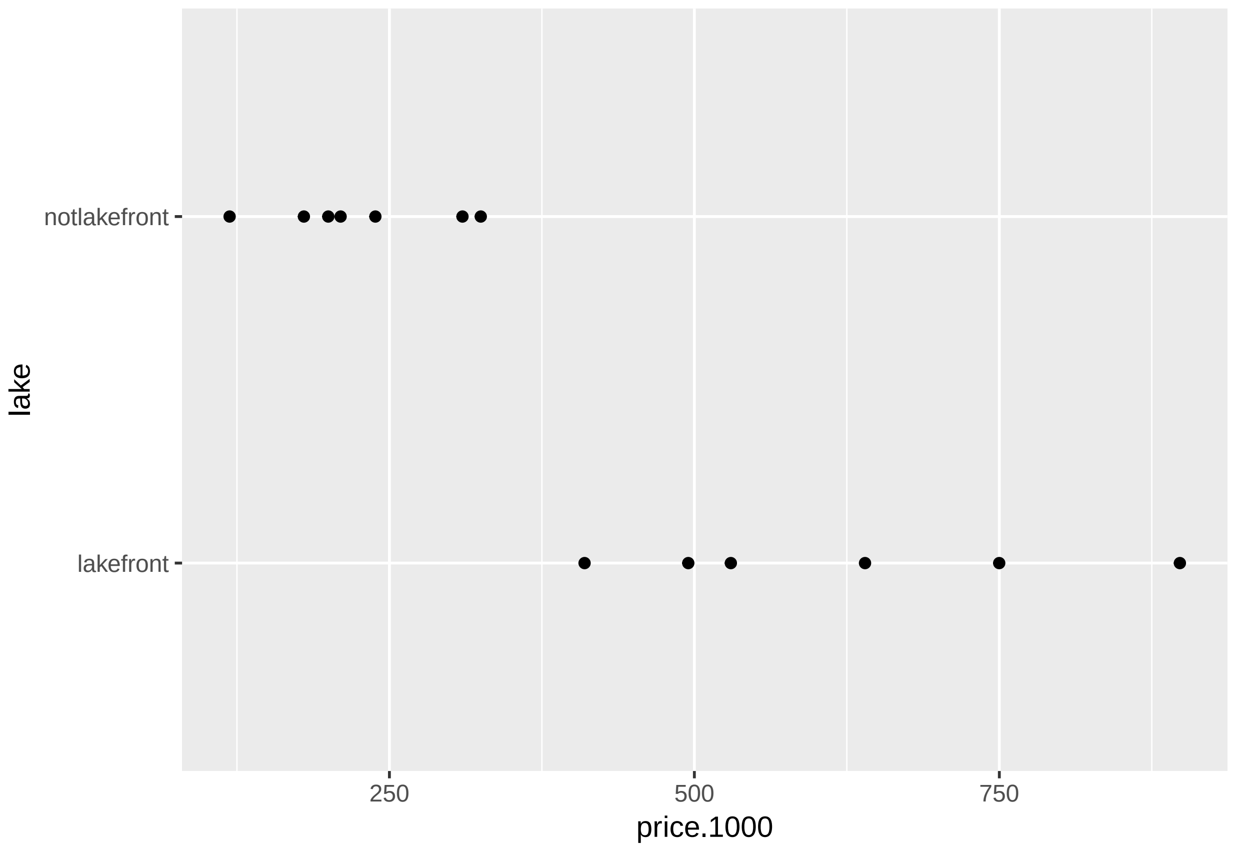
SOLUTION: Yes, we go from 185.1 thousand dollars in the first model to 49.48 thousands dollars in the new model

16. Based on the scatterplot, is location confounded with size of the home?

```
home.data %>%
  ggplot(aes(x=sqft,y=lake)) +
  geom_point()
```



```
home.data %>%
  ggplot(aes(x=price.1000,y=lake)) +
  geom_point()
```



SOLTUION: Location DOES NOT appear to be confounded with home size; however, there does appear to be a relationship between location and price of the home (based on both scatter plots).

SOURCES OF VARIATION DIAGRAM: OBSERVED VARIATION IN: Home Prices INCLUSION CRITERIA: Homes north of Lake Macatawa, Michigan in 2015 SOURCES OF EXPLAINED VARIATION: Home Size (square footage), Location (Lake Front or Not Lake Front) SOURCES OF UNEXPLAINED VARIATION: Number of bed/baths, plot size, age of home

Step 4: Draw Inferences beyond the data

Coming in later chapters

Step 5: Formulate conclusions

17. Summarize the conclusions of the study, including the statistical model I would recommend. Is there a larger population of homes you are willing to generalize? Is it reasonable to conclude that either the size of the home and/or location of the property is causing variation in home prices?

SOLUTION: I would summarize by using the statistical model that includes location and home size, as this model explained more variance than the other models considered in the study. My conclusion is that location and home size has a positive impact on the home price - lake front properties are generally going to be more expensive and the larger the home on the lakefront will increase the property value. I would not be willing to generalize this study to a larger population of homes, as there are many other factors that could explain variability in home prices, such as school districts, location to major cities, etc. In this study, home size and location are extremely likely to cause variation in home prices.

18. Identify any limitation you see to this study. What additional data would you like to collect to answer this research question?

SOLUTION: Information about homes in other areas would be particularly useful to this study, as this study only accounts for two possible explanations of variations and is very specific to the Lake Macatawa area.