Exploration P.B: Housing Prices in Michigan

1. To predict home prices, we might look at location, square footage, number of bathrooms, and number of bedrooms, among many other possibilities.

2. Observational units: homes for sale just north of Lake Macatawa, Michigan in 2015. Response variable: sale price (in thousands of dollars) as listed on Zillow. This is quantitative.

   We would look at a histogram or boxplot to explore the data visually because the response is quantitative. We would use the mean or median to explore the typical price; and range, or standard deviation, or max and min to explore the variability of the sale prices.

3. The distribution of home sale prices in this sample is skewed to the right. There are more values below the mean than above it.

4. The mean would not be a reasonable value to estimate a typical home price because the distribution is not symmetric. Because the data are skewed to the right, the large values would inflate the mean. Instead, we look at the median.

5. The data are skewed to the right, so the mean will be larger than the median. The mean is pulled up by the large values and will “follow the tail.”

6. The standard deviation of these home prices is $240,920. Roughly, the typical distance of a home is $240,920 away from the mean. This will not provide an accurate prediction of home prices. Some possible explanations for this variation in home prices are size and location. We might expect the price of a home in the same location to cost more as size increases, and homes in a desirable location, such as close to a lake, may cost more.

7. Observed - predicted (mean) = 639,000 - 408,060 = 230,940.
   We would have been $230,940 below the actual price.

8. We under-predicted the price of the first house, so the residual is positive.

9. There are eight negative and six positive residuals, spread relatively equally around 0.

10. There is a moderate positive linear association between home price and square footage. A one-thousand square foot increase in size is associated with a price increase of approximately $200,000. This is what we predicted.

11.
    a. Predicted price = -59.37 + 0.2127 × sqft (in thousands of dollars), standard error of residuals = 185.073 thousand dollars

b. The intercept, -$59,370 (rounded), is the predicted home price when the square footage is zero. The slope of $212.7 means that each one square foot increase in size predicts an increase of $212.7 in sale price.
c. Predicted price = -59.37 + .2127*(2700) = $514,920
Observed - Predicted = 639,000 - 514,920 = $124,080
This residual is smaller than before. Using the mean as the predicted value, the residual was $230,940.
d. The standard error of the residuals for the mean housing price is $240,920 and for the linear model with square footage as a predictor is $185,073. It is lower for the linear model, indicating that the model accounting for square footage is better on average at predicting home prices.

12. Not lakefront: Predicted price = 58.11 + 0.0839 × sqft
Lakefront: Predicted price = 86.76 + 0.2199 × sqft

The intercept for lakefront homes ($86,760) is greater than for non-lakefront homes ($58,110). A zero square foot lakefront home is predicted to be more expensive than a zero square foot not lakefront home. This doesn't really make sense in context because a house cannot be zero square feet. Maybe if we consider an empty plot of land, it makes sense.

The slope for lakefront homes ($219.9) is greater than for not lakefront homes ($83.9). This means that a one square foot increase in size predicts a larger increase in price if the home is a lakefront property compared to not lakefront. This makes sense in context because we might expect lakefront homes to increase in price at a faster rate as size increases.

13. Predicted price for first home (Lakefront): 86.76 + 0.2199(2700) = $680,490
14. Standard Error of residuals: 49.479 thousand dollars

15. Including location further reduces the standard error of the residuals from 185.073 to 49.479 thousand dollars.

16. Location does appear to be related to price (lakefront homes are more expensive than not lakefront homes, on average), but there is only a weak association between location and size (lakefront homes tend to be larger).

Observed Variation in: Home sale prices	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria Homes north of Lake Macatawa, Michigan, in	• Size (square footage) • Location (lakefront or not)	• Individual variation from home to home • Age of house

Observed Variation in: Home sale prices	**Sources of explained variation**	**Sources of unexplained variation**
2015		• Number of bedrooms/bathrooms • Unknown

17. We would recommend the model with square footage and location. This model has the lowest residual standard error by a substantial amount (about 140 thousand dollars less than the model with only square footage). This reduction in residual standard error tells us that the variables size and location are accounting for variation in home prices. But we should leap to any “causal” conclusions from this observational study, there could be other variables that also change with size and location and that contribute to the home price (e.g., age). We can only apply these results to homes just north of Lake Macatawa, MI, and maybe just this time frame.

18. Our scope of inference is also limited (observational study, specific location). We could look into collecting information on homes in other areas to be able to make broader conclusions.