MA256 Lesson 18 & 19 - Multiple Variables and Interactions (Causality and Multiple Regression Supplement Ch3 & 4)

Many people say that you can predict the cost of a home (in thousands of USD) based on its size (square footage) because bigger homes cost more to build, so I'd like for you to verify this claim using a random sample of homes around Lake Michigan (Houses.csv).

A1.) Identify and classify the variables of interest.

Explanatory: Home size (in square feet) - Quantitative

Response: Price (in thousands of USD) - Quantitative

A2.) Define the coefficient symbols and interpret what they represent (in context) given the simple linear regression equation in the form of $y = \beta_0 + \beta_1 * sqft$.

β_0 (Intercept Coefficient): the average price of a 0 sqft home (vacant lot).

β_1 (Slope Coefficient): the population slope is the predicted increase in average price (in thousands of USD) for every one-square-foot increase in home size.

3. Express the null and alternate hypotheses in words & symbols.

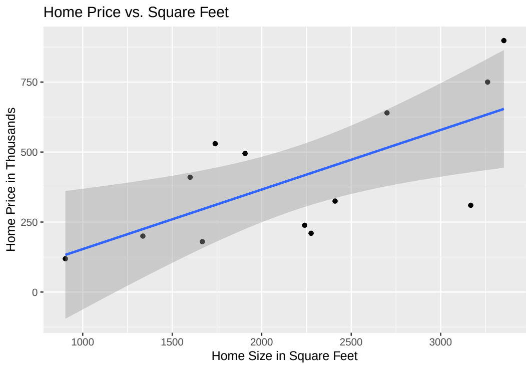
$H_0 : \beta_1 = 0$. There is no association between the size of a home and its price.

$H_a : \beta_1 > 0$. There is a positive association between the size of a home and its price.

4. Create a scatterplot of Home Size vs Price, and add a layer (`geom_smooth`) to include the best-fitting least squares regression line. Are the 4x validity conditions met for us to use the theory-based approach to assess strength of evidence? Use headers to explicitly state each of them, and then clarify if they are met.

```
houses %>% ggplot(aes(x = sqft, y = `price in thousands`)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x="Home Size in Square Feet", y = "Home Price in Thousands",
       title = "Home Price vs. Square Feet")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Linearity: The data seems to generally follow a linear trend.

Independence (New): It makes sense the home prices are independent from one another e.g. if my neighbor's house has ugly paint inside that won't influence my house price – or – if my neighbor's house only has 2 bathrooms but mine has 4 that won't affect my home price.

Normality: Equal number of observations above (6) and below (7) the line.

Equal Variance: There is an approximately equal variance throughout the entire domain, given flexibility due to small sample size.

5. Use the Course Guide example R Code for Lesson 18-19 as a template to create Coefficient Table and then write the complete equation for House_Model given the form of $y = \beta_0 + \beta_1 * sqft$.

```
house.lm <- houses %>% lm(`price in thousands` ~ sqft, data = .)
summary(house.lm)
```

```
##
## Call:
## lm(formula = `price in thousands` ~ sqft, 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
```

Price= -59.36870 + 0.21274*sqft

6. Interpret your slope and intercept coefficients:

Slope Coefficient: The increase in average price (in thousands of USD) for every 1 sqft increase in home size is 0.21274. Alternative Interpretation: \sol{The increase in average price for every 1 square foot increase in home size is $212.74 (0.21274 * 1,000 since price is in thousands).}

Intercept Coefficient: The average price (in thousands of USD) of a 0 square foot home is -59.36870. Alternative Interpretation: \sol{The average price of a 0 square foot home is -$59,368.70 (-59.36870 * 1,000 since price is in thousands).}

7. Would you be comfortable using your model to predict the price of a 0 sqft home (vacant lot)? Why or why not?

\sol{Since positive numbers indicate money flowing from a buyer to a seller to purchase their property, and since the intercept coefficient is negative, it doesn't make sense that someone would pay me $59,368.70 to "purchase" a vacant lot i.e. 0 square foot home. I am not comfortable using my model in this way. WE MUST ALWAYS USE CAUTION WHEN EXTRAPOLATING WITH OUR MODEL. }

Bonus: If we want to predict prices for vacant lots, what should we use to create a model?

If we wanted to predict the prices of vacant lots, perhaps we should only use vacant lots to create a new least squares regression model.

8. Calculate the 95% confidence interval for the slope statistic and provide an interpretation of the CI.

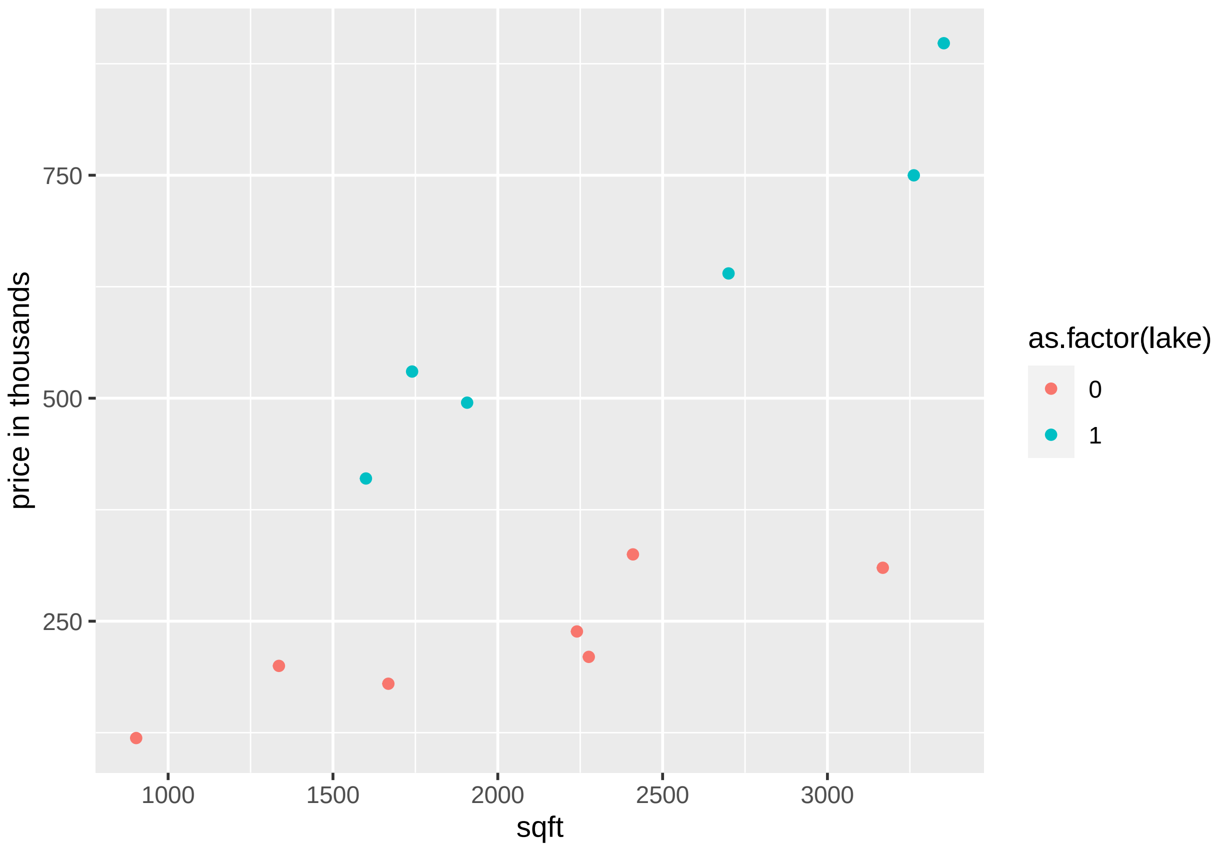
\sol{The 2SD 95% CI = Statistic ± 2*SD. Since the slope coefficient (statistic) is 0.21274, and since its Standard Error (SE) is 0.06963, the 95% CI can be calculated as follows:\ Lower Bound: 0.21274 – (2*0.06963) = 0.07348\ Upper Bound: 0.21274 + (2*0.06963) = 0.352\ 95% CI: (0.07348, 0.352)\ I am 95% confident that a 1 sqft increase in home size is associated with an average increase of 0.07348 thousands of USD to 0.352 thousands of USD in home price for homes around Lake Michigan.}

9. Briefly explain why location i.e. lakefront (=1) or non-lakefront (=0), may be another explanatory variable.

Lake houses are traditionally more expensive than non-lakefront houses. Lake impacts the response variables.

10. Does lakefront affect price and sqft? Create a new scatterplot that is colored based on lakefront status.

```
houses %>% ggplot(aes(x = sqft, y = `price in thousands`, color = as.factor(lake))) +
  geom_point()
```



Based on the scatterplot results, it seems like even the smallest lakefront homes are more expensive than the largest non-lakefront homes that are twice the size (in sqft).

11. We can adjust for the additional explanatory variable by including it in a multiple regression model. In this model, we estimate the price per square foot when we hold the location of the house constant.

$price_i = \alpha_0 + \alpha_1 * sqft_i + \alpha_2 * lake_i$

where $price_i$ is the price of house i, $sqft_i$ is the size (sq ft) of house i, and $lake_i$ is 1 if house i is lakefront and 0 if not lakefront. It may help to think of "i" as a counter from house #1 to #13

Create a new model with lake as an additional variable & use the example code below to create a scatter plot to view the new model.