**DATA 557**

**Winter 2022**

**Homework Assignment 4**

**Instructions**

Submit your solutions **in pdf format** to the dropbox on the canvas page by **12:00PM, Friday February 18**. You may use any program to generate your pdf file. (RStudio is recommended but not required.)

For each question you will be given 1 point for complete credit, ½ point for partial credit, and 0 points for no credit. Assignment of credit will be based on the correctness of your answers as well as your reasoning (when requested as part of the question). You do not need to submit R code for this assignment except where it is requested.

You may work together to help each other solve problems, but you should create your own solutions and hand in your own work without copying others’ work.

**Data: ‘Sales.csv’**

The data consist of sales prices for a sample of homes from a US city and some features of the houses.

Variables:

LAST\_SALE\_PRICE: the sale price of the home

SQFT: area of the house (sq. ft.)

LOT\_SIZE: area of the lot (sq. ft.)

BEDS: number of bedrooms

BATHS: number of bathrooms

1. Calculate all pairwise correlations between all five variables.

2. Make a scatterplot of the sale price versus the area of the house. Describe the association between these two variables.

3. Fit a simple linear regression model (Model 1) with sale price as response variable and area of the house (SQFT) as predictor variable. State the estimated value of the intercept and the estimated coefficient for the area variable.

4. Write the equation that describes the relationship between the mean sale price and SQFT.

5. State the interpretation in words of the estimated intercept.

6. State the interpretation in words of the estimated coefficient for the area variable.

7. Add the LOT\_SIZE variable to the linear regression model (Model 2). How did the estimated coefficient for the SQFT variable change?

8. State the interpretation of the coefficient of SQFT in Model 2.

9. Report the R-squared values from the two models. Explain why they are different.

10. Report the estimates of the error variances from the two models. Explain why they are different.

11. State the interpretation of the estimated error variance for Model 2.

12. Test the null hypothesis that the coefficient of the SQFT variable in Model 2 is equal to 0. (Assume that the assumptions required for the test are met.)

13. Test the null hypothesis that the coefficients of both the SQFT and LOT\_SIZE variables are equal to 0. Report the test statistic.

14. What is the distribution of the test statistic under the null hypothesis (assuming model assumptions are met)?

15. Report the p-value for the test in Q13.