**DATA 557**

**Winter 2022**

**Homework Assignment 5**

**Instructions**

Submit your solutions **in pdf format** to the dropbox on the canvas page by **1:00PM, Friday February 25**. 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\_sample.csv”.**

The data are a random sample of size 1000 from the “Sales” data (after removing observations with missing values).

1.1. Fit a linear regression model (Model 1) with sale price as response variable and SQFT, LOT\_SIZE, BEDS, and BATHS as predictor variables. Add the fitted values and the residuals from the models as new variables in your data set. Show the R code you used for this question.

1.2. Create a histogram of the residuals. Based on this graph does the normality assumption hold?

Answer the following questions using residual plots for the model. You may make the plots using the residuals and fitted variables added to your data set or you may use the ‘plot’ function. You do not need to display the plots in your submission.

1.3. Assess the linearity assumption of the regression model. Explain by describing a pattern in one or more residual plots.

1.4. Assess the constant variance assumption of the regression model. Explain by describing a pattern in one or more residual plots.

1.5. Assess the normality assumption of the linear regression model. Explain by describing a pattern in one or more residual plots.

1.6. Give an overall assessment of how well the assumptions hold for the regression model.

1.7. Would statistical inferences based on this model be valid? Explain.

1.8. Create a new variable (I will call it LOG\_PRICE) which is calculated as the log-transformation of the sale price variable. Use base-10 logarithms. Fit a linear regression model (Model 2) with LOG\_PRICE as response variable and SQFT, LOT\_SIZE, BEDS, and BATHS as predictor variables. Report the table of coefficient estimates with standard errors and p-values.

1.9. Give an interpretation of the estimated coefficient of the variable SQFT in Model 2.

Answer the following questions using residual plots for Model 2. You do not need to display the plots in your submission.

1.10. Assess the linearity assumption of Model 2. Explain by describing a pattern in one or more residual plots.

1.11. Assess the constant variance assumption of Model 2. Explain by describing a pattern in one or more residual plots.

1.12. Assess the normality assumption of Model 2. Explain by describing a pattern in one or more residual plots.

1.13. Give an overall assessment of how well the assumptions hold for Model 2.

1.14. Would statistical inferences based on Model 2 be valid? Explain.