Predictive Analytics Assignment
worth 20% of your final grade.
Due the 17th November at 23:59pm

The data file "House.csv" contains information on the sale of 76 single-family homes in Dublin during 2005. We will model single-family home sale price by (Price in excel file $= \frac{\text{Price the house sold for in thousands of eur}}{1000}$), which range from 155.5 (€155,500) to 450.0 (€450,000), using these predictor variables:

Size floor size (thousands of square feet)

Lot lot size category (from 1 to 11 explained below)

Bath number of bathrooms (with half-bathrooms counting as 0.1 explained below)

Bed number of bedrooms (between 2 and 6)

Year year the house was built.

Garage garage size (0, 1, 2, or 3 cars)

High indicator for The High School (reference: The High School)

Alexandra indicator for Alexandra College (reference: Alexandra College)

Stratford indicator for Stratford College (reference: Stratford College)

St.Mary's indicator for St.Mary's College (reference: St.Mary's College)

St Louis indicator for St Louis High School (reference: St Louis High School)

It seems reasonable to expect that homes built on properties with a large amount of land area command higher sale prices than homes with less land, all else being equal. However, an increase in land area of (say) 2000 square feet from 4000 to 6000 should probably make a larger difference (to sale price) than going from 24,000 to 26,000. Thus, realtors have constructed lot size "categories," which in their experience correspond to approximately equal-sized increases in sale price.

Lot Size	0-3000ft^2	3000-5000ft^2	5000-7000ft^2	7000-10,000ft^2
Category	1	2	3	4
Lot Size	10,000-15,000ft^2	15,000-20,000ft^2	20,000ft^2-1 acre	1-3 acres
Category	5	6	7	8
Lot Size	3-5 acres	5-10 acres	10-20 acres	
Category	9	10	11	

To reflect the belief that half-bathrooms (i.e., those without a shower or bathtub) are not valued by home-buyers nearly as highly as full bathrooms, the variable Bath records half-bathrooms with the value 0.1.

Instructions: Please provide your R code and a written report. Answer the following questions in your written report (use single spacing and a font size 12). There is a word limit of 500 words for each question. Please write your name and student id at the top of your r code and your written report.

Exploratory Data Analysis:

1. Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.
2. Convert all the categorical variables to factors. Using the summary and a boxplot describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.
3. Using the summary, correlation and the pairs plots discuss the relationship between the response sales price and each of the numeric predictor variables.

Regression Model:

1. Fit a multiple linear regression model to the data with sales price as the response and size, lot, bath, bed, year, garage and school as the predictor variables. Write down the equation for this model.
2. Interpret the estimate of the intercept term β_0.
3. Interpret the estimate of $\beta_{\texttt{size}}$ the parameter associated with floor size (`Size`).
4. Interpret the estimate of $\beta_{\texttt{Bath1.1}}$ the parameter associated with one and a half bathrooms.
5. Discuss and interpret the effect the predictor variable bed on the expected value of the house prices.
6. List the predictor variables that are significantly contributing to the expected value of the house prices

7. For each predictor variable what is the value that will lead to the largest expected value of the house prices.

8. For each predictor variable what is the value that will lead to the lowest expected value of the house prices.

9. By looking at the information about the residuals in the summary and by plotting the residuals do you think this is a good model of the expected value of the house prices.

10. Interpret the Adjusted R-squared value.

11. Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

ANOVA:

1. Compute the type 1 anova table. Interpret the output. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

2. Which predictor variable does the type 1 anova table suggest you should remove the regression analysis.

3. Compute a type 2 anova table comparing the full model with all predictor variables to the the reduced model with the suggested predictor variable identified in the previous question removed. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

Diagnostics:

1. Check the linearity assumption by interpreting the added variable plots and component-plus-residual plots. What effect would non-linearity have on the regression model and how might you correct or improve the model in the presence of non-linearity?

2. Check the random/i.i.d. sample assumption by carefully reading the data description and computing the Durbin Watson test (state the hypothesis of the test, the test statistic and p-value and the conclusion in the context of the problem). What are the two common violations of the random/i.i.d. sample assumption? What effect would dependant samples have on the regression model and how might you correct or improve the model in the presence of dependant samples?

3. Check the collinearity assumption by interpreting the correlation and variance inflation factors. What effect would multicollinearity have on the regression model and how might you correct or improve the model in the presence of multicollinearity.

4. Check the zero conditional mean and homoscedasticity assumption by interpreting the studentized residuals vrs fitted values plots and the studentized residuals vrs predictor variable plots. What effect would heteroscedasticity have on the regression model and how might you correct or improve the model in the presence of heteroscedasticity.

5. Check the Normality assumption by interpreting the histogram and quantile-quantile plot of the studentized residuals. What effect would non-normality have on the regression model and how might you correct or improve the model in the presence of non-normality.

Leverage, Influence and Outliers:

1. What is a leverage point? What effect would a leverage point have on the regression model? Use the leverage values and the leverage plots to see if there is any leverage points.

2. What is an influential point? What effect would an influential point have on the regression model? Use the influence plot to see if there is any influence points.

3. What is an outlier? What effect would an outlier have on the regression model? How would you correct for outliers? Use the outlier test and outlier and leverage diagnostics plot to see if there is any outliers. Deal with the outliers if any are identified.

Expected Value, CI and PI:

1. Plot the observed house prices, their expected vale (fitted value), confidence intervals (in red) and prediction intervals (in blue). Looking at this plot is this model providing a good estimate of the house prices.