Task: Investigate factors affecting house prices in the USA (Baton Rouge, Louisiana). You will begin with an exploratory data analysis. You will then answer some specific questions about your data set before fitting simple and multiple linear regression models. Note that the data set contains a subset of the variable *style*.

For the PDF file, do not print out the code chunks, only the text i.e., your analysis and any desired output from the code chunks. Ensure that it is well presented. For example, plots should include appropriate axis labels, titles and/or captions. Text should be clear and without spelling errors.

1. Data exploration Summarise your data set and carry out an exploratory data analysis. You may want to include the following:
   * A brief summary of the variables in your data set. For example, how many variables and observa- tions are there in your data set? What are the different variable types? For qualitative variables, what are the categories?
   * Descriptive statistics. Consider including either the trimmed or Winsorized mean, if appropriate. Present your results clearly in a table and comment on your findings.
   * Appropriate plots visualising the distributions of some of the variables. Comment on your findings.
   * Correlations between variables. Comment on your findings.

This section should not be excessively long. Aim to highlight/summarise key aspects of your data set. Be concise and clear.

1. Probability, probability distributions and confidence intervals.Calculate ALL of the following:
   * A house is chosen at random from your data set. Compute:  
     **–** the probability it has a pool.  
     **–** the (conditional) probability it has a fireplace, given that it has a pool.
   * Find the probability that, out of 10 houses chosen at random from your data set, at least 3 will have a pool. Explain your working.
   * Assuming the data set provides a random sample of houses in the USA, calculate a 95% confidence interval on the mean house price in the USA.
2. Contingency tables and hypothesis tests**.** Carry out ALL the following tasks:
   * Test the hypothesis that the mean house price (over all house styles) is greater if a house is on the waterfront. Use a 5% significance level. Clearly state the test used and the null and alternative hypotheses. Interpret the result.
   * Create a contingency table showing relative frequencies for "Pool" and "No pool" according to whether a house has or hasn’t got a fireplace. Print the table out neatly in your report, including row and column names.
   * Using a 5% significance level, test whether a house having a fireplace is independent of whether it has a pool. Comment on your results.
3. Simple Linear Regression. Address ALL of the following:
   * Perform a simple linear regression with ln(*price*) as the response variable and ln(*sqft*) as the predictor variable. Provide the fitted model. Determine from your output if total area is a significant predictor of house price. Interpret the slope. NOTE: ln() is the natural logarithm.
   * Show on a scatter plot both the data and the fitted model. On a separate plot, show the residuals between the model and the data, including a reference line for zero residuals. Include axis labels. Comment on the plots.

5. Multiple Linear Regression. Carry out ALL the following tasks:

* Perform a multiple linear regression of ln(*price*) against all the predictor variables (refer to this as the "full model"). Use ln(*sqft*) and deal appropriately with qualitative variables. Explain clearly what you have done. Provide and comment on the fitted model.
* Use feature selection to produce a reduced model. Justify your method.
* Using k-fold cross validation, comment on whether the reduced model obtained following feature selection performs better than the full model. Use an appropriate metric to compare the model performances.