Instructions: You must submit this homework by pushing the "hw2.Rmd" file to your team's repo. Note that that is the **only** file you will be allowed to push. Commit early and often.

Research Scenario: Predicting house prices People buying or selling houses would like to know how much they can expect to get, or pay, for a property. This is also a concern for those who are making mortgage loans, or for those taxing real estate (and who are more likely to commission statistical studies than individual home-owners). The price of a house depends on its physical characteristics, including size, features, quality of construction, age, etc. It also depends on location, and current market characteristics. You are approached by a research group which has data on a sample of residential sales in a midwestern city; the variables are described in the table below. They would like you to fit a multiple linear regression, with sales price as the response and (some or all of) the other variables as predictors.

Variable name	Description
Sales price	Sale price of house (dollars)
Finished square feet	Finished area of residence(square feet)
Number of bedrooms	Total number of bedrooms in residence
Number of bathrooms	Total number of bathrooms in residence
Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
Garage size	Number of cars that garage will hold
Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
Year built	Year property was originally constructed
House Quality	1= high quality, 2 = medium, 3 = low
Lot size	Lot size (square feet)
Adjacent to highway	1 if the property is adjacent to a highway, 0 otherwise

Your client believes that higher quality of construction should predict higher prices. They also believe that older houses tend to have lower prices, though this relationship is thought to differ depending on whether or not the house is adjacent to a highway. They also think that the relationship between price and finished area differs depending on the number of bedrooms.

Your assignment

You are to begin analysis of this dataset in order to try to answer the client's questions. Your goal is to communicate statistical conclusions clearly both using appropriately chosen words and graphics.

The `hw2.Rmd` file in your team's repo is separated into 3 main sections: (1) Introduction, (2) Exploratory data analysis, and (3) Initial modelling. Each section has a number of points you should discuss, suggests graphics you should create, and gives some code which may help to do certain tasks. You should address all of the points in some way or another. This may mean that you have to write a few paragraphs and produce multiple figures for some points. Others you may consider to be ill-advised given previous work. In such a case, you should say so and indicate why.

Writing advice

Your language should be very clear and precise. Do not make claims for which you have no evidence. Do not say "will" or "would" when you really mean "may" or "might". Do not use language that implies causation; you are studying associations between variables only. Move away from wordy phrases (e.g: This is because,

this is due to, the reason that this is, this means that, I believe that this, I think the reason is that). Make sure pronouns have clear referents ("these results show" vs. "this shows").

Account for your audience. Interpret results in the context of the problem. A statement like "The p-value is less than 0.05, so we reject the null hypothesis" will receive no credit. Instead say, "To determine whether the number of house cats is a good predictor of hermit-ness, we examine the regression coefficient. We see that this coefficient is very large, suggesting that even a 1-cat increase predicts that the owner is more likely to be a hermit. Furthermore, the associated p-value is very small indicating..."

Grading rubric

Words (4) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (2) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (4) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text or referred to with convenient labels.

Code (10) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. The text of the report is free of intrusive blocks of code. With regards to R Markdown, all calculations are actually done in the file as it knits, and only relevant results are shown.

Analysis (10) Variables are examined individually and bivariately. Features/observations are discussed with appropriate figure or tables. The relevance of the EDA to the modeling is clearly explained. The model's formulation is clearly related to the substantive questions of interest. The model's assumptions are checked by means of appropriate diagnostic plots or formal tests; if the model is re-formulated, the changes are both well-motivated by the diagnostics, and still allow the model to answer the original substantive question. Limitations from un-fixable problems are clearly noted. The substantive questions about real estate pricing are answered as precisely as the data and the model allow. The chain of reasoning from estimation results about models, or derived quantities, to substantive con- clusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the discussion.

Extra credit (5) Up to five points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.