Introduction

Appraising residential real estate — predicting the price at which it could be sold, under current market conditions — is important not only for people buying and selling houses to live in, but also for real estate developers, mortgage lenders, and local tax assessors. Currently, appraisal is usually done by skilled professionals who make a good living at it, so naturally there is interest in replacing them by machines. In this report, we investigate the feasibility of real estate appraisal by means of linear statistical models.

Specific points of interest to the client include the relationship between the quality of the house's construction and its price; the relationship between age and price, and whether this changes depending on proximity to a highway; and the relationship between price, the finished area of the house, and the number of bedrooms.

Exploratory data analysis

The data, supplied by an undisclosed client, come from a selection of "arms-length" residential real estate transactions in an unnamed city in the American midwest in 2002. This records, for 522 transactions, the sale price of the house, its finished area and the area of the lot, the number of bedrooms, the number of bathrooms, the number of cars that will fit in its garage, the year it was built, whether it has air conditioning, whether it has a pool, whether it is adjacent to a highway, and the quality of construction, graded from low to medium or high. It is notable that, except for highway adjacency, we have no information about the location of the houses, though this is proverbially a very important influence on their price, through access to schools, commuting time, land value, etc.

Pairwise scatter-plots for the quantitative variables (Figure 1) show that, unsurprisingly, there is a positive relationship between price and area (stronger for finished area than the total lot size), and price and the number of bedrooms, bathrooms, or garage slots (all three of which are strongly positively related to each other). The relation between price and these three "count" variables could well be linear. There is a positive relation between price and the year of construction, i.e., newer houses cost more. Newer houses also tend to be larger, both in finished area and the number of rooms, though not to have bigger lots.

Inspection of the plots shows there is one record with 0 bedrooms, 0 bathrooms, and a three-car garage with air conditioning. This is either not a piece of residential real estate, or its data is hopelessly corrupt; either way, we drop it from the data from now on.

Box-plots, showing the conditional distribution of price for each level of the categorical predictors, suggest that houses with air-conditioning and pools are more expensive, that being next to a highway makes little difference, and that higher quality of construction implies, on average, higher prices. The mid-points of the boxes for quality don't *quite* fall on a straight line, so treating quality as a numerical variable isn't obviously compelling, but not clearly crazy either.

Initial Modeling

To answer the client's questions, our model should include quality, finished area, the number of bedrooms (and the interaction between those two), and the year the house was built and whether it is adjacent to a highway (and the interaction between those two). Based on our EDA, it also seems reasonable to include air-conditioning and pools. We deliberately left out the number of bathrooms, the size of the garage, and the size of the lot. While price seems to be linearly related to the number of bedrooms, we include it as a factor, both to check that, and to get three distinct slopes for price on finished area as quality varies.

This initial model has a root-mean-squared error of \$ $\pm 5.94 \times 10^4$, which is not shabby when the median house price is \$ 2.3×10^5. Before passing to issues of model selection, however, such as whether all the interactions are necessary, whether discrete variables might be usefully recoded, etc., let's look at the diagnostic plots.

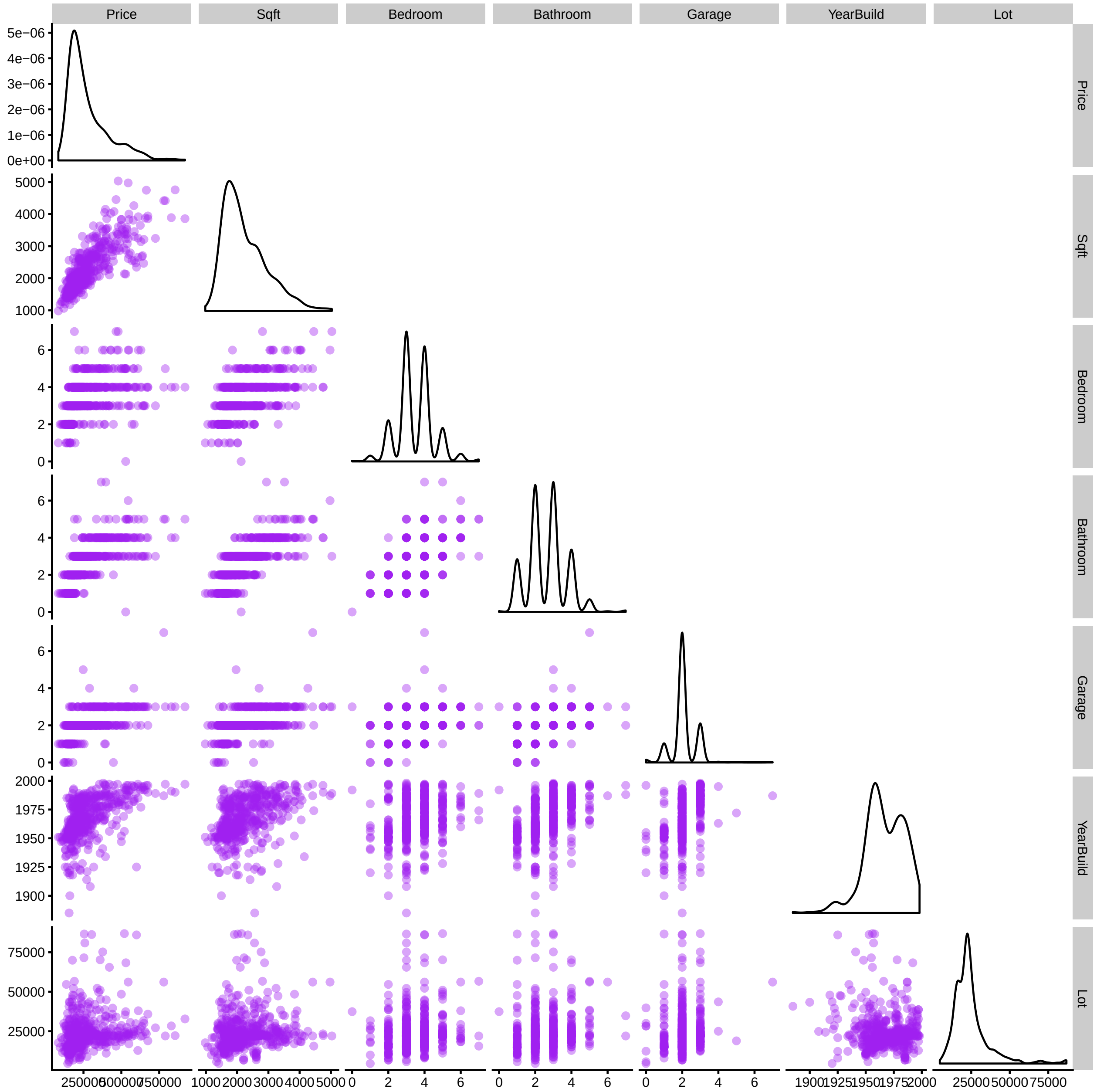

Figure 1: *Pairs plot for quantitative variables*

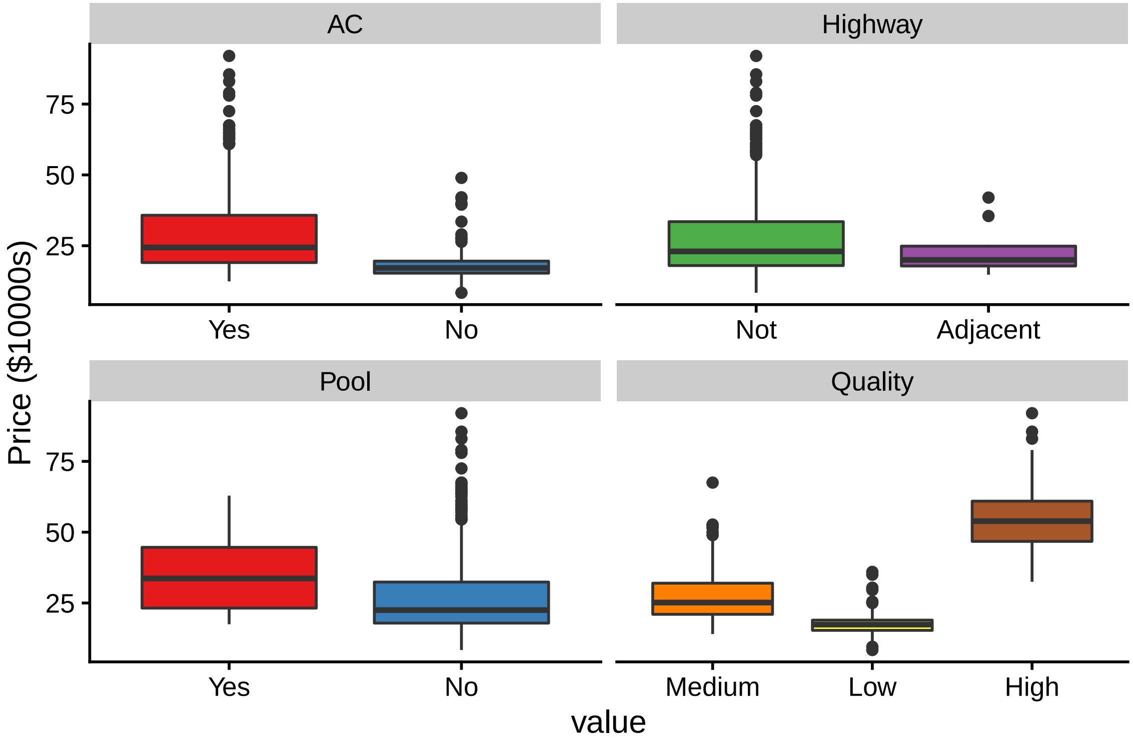

FIGURE 2: *Conditional distributions of price given qualitative predictors. Box widths reflect the number of points in each group, notches show medians plus/minus a margin of error.*

The first thing to say is that the distribution of the residuals doesn't look very Gaussian, and a Box-Cox transformation suggests the un-intuitive, indeed un-interpretable, transformation $1/\sqrt[3]{Y}$.

Clients who ask for a model of prices are rarely happy with models for the inverse cubic roots of prices, so we must be doing something wrong. Examining plots of residuals versus predictors suggests that lot size matters after all, at least for big lots. The plots also suggest that houses built after ≈ 1980 are worth more than the model anticipates. The distributions of residuals conditional on discrete predictors, however, actually look mostly homogeneous.

Outliers

In addition to the house with no bedrooms or bathrooms, examination of Cook's distance shows two houses with exceptional influence over the model.

On examination, these are quite weird: small in area, fairly cheap, but heavy on bedrooms. These look more like rental properties than residences. Checking the pairs plot again shows no other such anomalies, so we delete them but leave the rest alone. Re-doing the other diagnostic plots shows little over-all change, however (figures omitted).

Price	Sqft	Bedroom	Bathroom
190000	2812	7	5
219900	1852	6	3

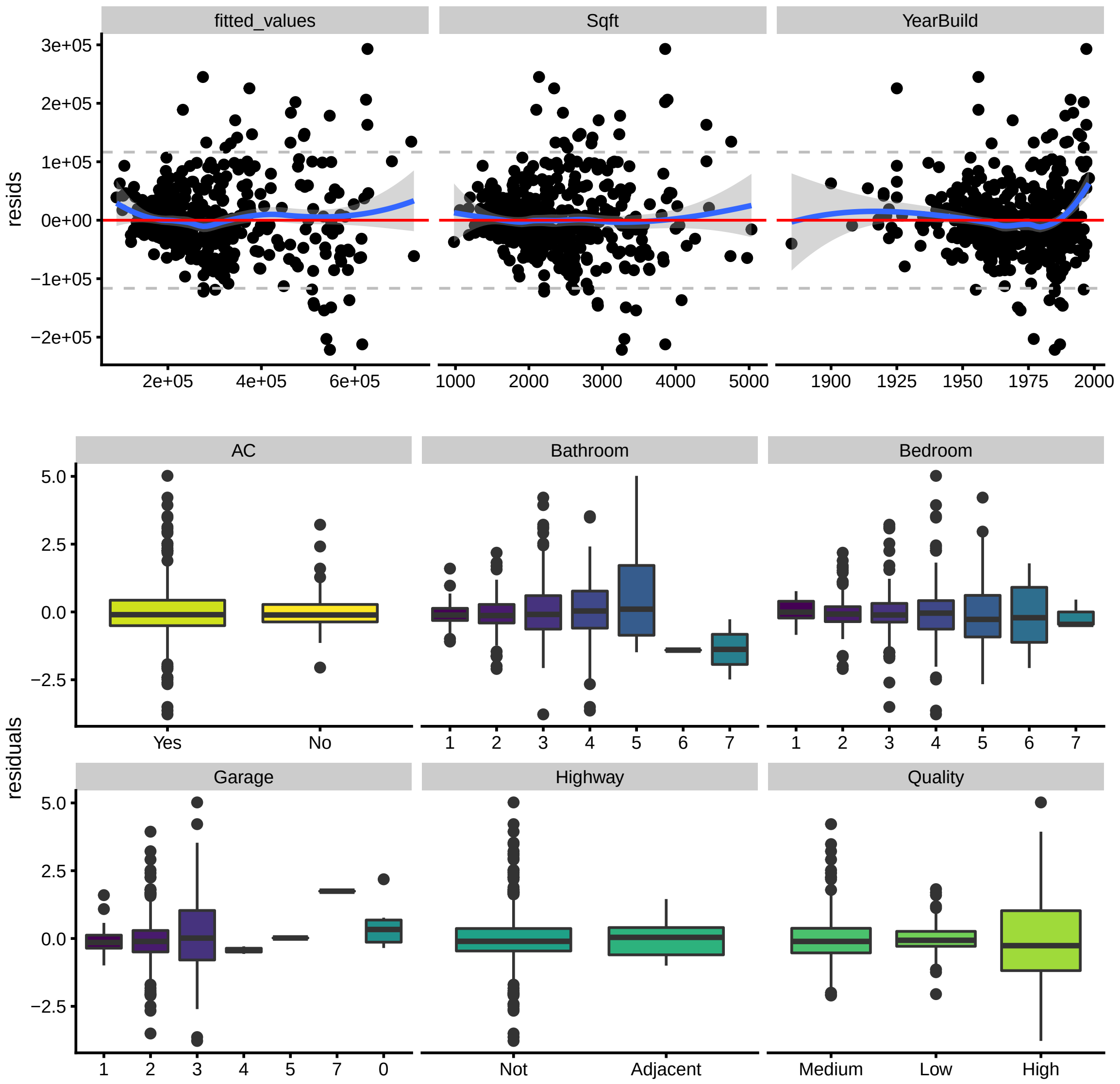

FIGURE 3: *Residuals versus fitted values and continuous predictors, and versus the discrete predictors. Grey lines are smoothing splines; dotted lines indicate plus/minus 2 standard deviations, either constant (red) or from a spline smoothing of the squared residuals (grey).*

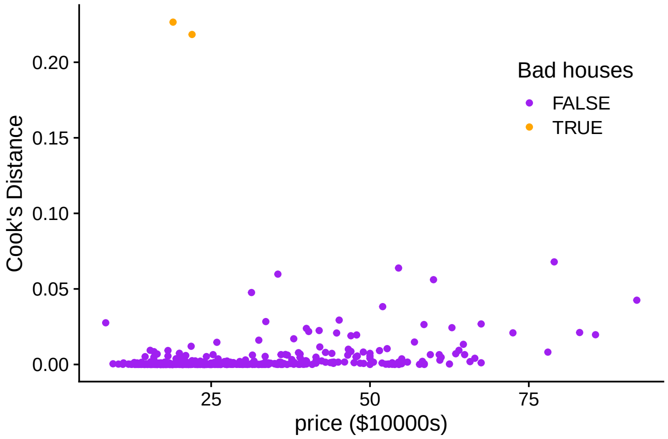

FIGURE 4: *Cook's distance for each data point: extremely influential points are flagged in red.*

Model selection / evaluation

Our examination of the data and the diagnostics plots suggest several possible modifications to our baseline model: adding a term for lot size, letting the slope on year change for younger houses, recoding bedrooms to the three levels, and turning the interactions of interest off. Each of these five choices is logically independent of the others, yielding $2^5 = 32$ possible models.

Rather than doing 32 sets of diagnostics, we will use leave-one-out cross-validation to select a model. That is, we'll fit each model on $n - 1$ of the data points, see how well they predict the n^{th} data point without having seen it, and average squared errors across data points for each model. This gives us a good estimate of how well the models would predict new data, which is ultimately what the client cares about.

Since we also want to do statistical inference on our selected model, however, we will run the cross-validation on only *half* of the data; the other half will be used just for inference on the final, selected model. If we used the same data twice, for both selection and for inference, we'd exaggerate the precision of our inferences, basically because we'd be asking how well our model fit the data it was selected to fit. Splitting the data by taking a random sample of half the data points and keeping it aside for inference avoids this problem.

Compared to our initial model, the best-predicting model replaces the giving each bedroom number its own contrast with the three-level coding of bedrooms; adds lot size as a predictor; and lets more-recent (post-1980) houses have their own slope on the year of construction. Everything else, including all the interactions, is the same. Our confidence that these are mostly good choices is reinforced by the fact that the second-best model makes all the same changes to the initial model, except that it drops the interaction between year of construction and adjacency to the highway.

Final model/inference

Having selected our model on one half of the data, we may now do statistical inference on the other half, and not *guarantee* that the results are invalid. Table 1 gives point estimates, standard errors, p-values, and 95% confidence intervals for all the coefficients. This is strongly suggestive of higher quality construction predicting higher prices, since both the contrast coefficients for lower quality are negative, by hundreds of thousands of dollars.

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	-1.47e+06	6.51e+05	-2.250	2.51e-02	-2.75e+06	-1.85e+05
QualityLow	-1.14e+04	9.44e+03	-1.210	2.29e-01	-3.00e+04	7.20e+03
QualityHigh	1.43e+05	1.30e+04	11.000	5.27e-23	1.17e+05	1.69e+05

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
Sqft	1.99e+01	1.01e+02	0.198	8.43e-01	-1.78e+02	2.18e+02
YearBuild	7.92e+02	3.10e+02	2.560	1.12e-02	1.82e+02	1.40e+03
AdjHighway1	-3.24e+04	2.31e+04	-1.400	1.62e-01	-7.78e+04	1.31e+04
Airconditioning0	-8.41e+03	1.03e+04	-0.820	4.13e-01	-2.86e+04	1.18e+04
PoolYes	1.88e+04	1.46e+04	1.290	1.99e-01	-9.91e+03	4.75e+04
BedsCoded2–4	-9.60e+04	1.83e+05	-0.523	6.01e-01	-4.57e+05	2.65e+05
BedsCoded5+	3.44e+04	1.90e+05	0.182	8.56e-01	-3.39e+05	4.08e+05
Since1980TRUE	-4.52e+06	2.80e+06	-1.610	1.09e-01	-1.00e+07	1.01e+06
Lot	1.73e+00	3.52e-01	4.920	1.56e-06	1.04e+00	2.43e+00
Sqft:BedsCoded2–4	7.68e+01	1.01e+02	0.763	4.46e-01	-1.21e+02	2.75e+02
Sqft:BedsCoded5+	3.31e+01	1.02e+02	0.325	7.45e-01	-1.67e+02	2.34e+02
YearBuild:Since1980TRUE	2.29e+03	1.41e+03	1.620	1.06e-01	-4.92e+02	5.07e+03

Table 1: Point estimates and inferential statistics for our selected model.

The client's other specific questions are best answered using figures. The slope plots (Figure 5, left), shows the contribution made to the predicted price by the finished area for each level of the number of bedrooms. This suggests that, first, more area predicts a higher price, but, second, at equal area, more bedrooms predicts a higher price, except *maybe* for the very largest houses. I qualify the conclusion this way because there are very few 2–4 bedroom houses over about 4000 square feet (and no one-bedroom houses), where the predicted line for 2–4 bedrooms goes above that of 5+ bedrooms.

Turning to the right panel of the figure, it shows younger houses are predicted to have higher prices, with the premium on youth turning up fairly sharply if the house was built after 1980.

Relying very much on the exact confidence intervals and p-values from Table 1 is a bit dubious. The model predicts pretty well: its leave-one-out RMSE is $\pm\$5.6 \times 10^4$, while that of our original model is $\pm\$6.18 \times 10^4$. Moreover, plotting residuals against fitted values shows almost exactly no trend to the former, and reasonably constant variance (Figure 6, left). (Plots of residuals against predictors [omitted] are similarly good.) But those same residuals are still not very Gaussian (Figure 6, right). Thus, while the inferential statistics tell us that none of the interaction terms are significantly different from zero, or even estimated to within closer than $\pm\$10^5$ (except the contrasts for the house having been built since 1980), we don't know how much trust we can put in those results. If only we knew a way of doing inference without assuming Gaussian noise...

Conclusion

We have found a model which predicts the price of houses to within about 5.6×10^4. The coefficients in this model all make sense after we see them [1]: high-quality construction predicts higher prices, as do amenities like pools and air-conditioning, as do bigger lots, more floor space, and more bedrooms. Younger houses command a premium, especially houses built since 1980. It is somewhat annoying that the residuals remain not-very-Gaussian, but that's because our statistical technique is still too weak to do inference under these conditions, not because the model is wrong.

Because age (year of building) emerged as one of the strongest predictors, it would be good to know how it's linked to price. Perhaps, since neighboring houses tend to be of similar ages, it's acting as a proxy for location. Indeed, probably the single biggest thing missing from this data set is location. Even without it, though, we can do a tolerable job of rolling our own Zillow, and could do better with more data.

[1] Of course, as the sociologist Duncan Watts says, "Everything is obvious, once you know the answer". His book of that title can hardly be too highly recommended to anyone who is going to spend time interpreting models like these.

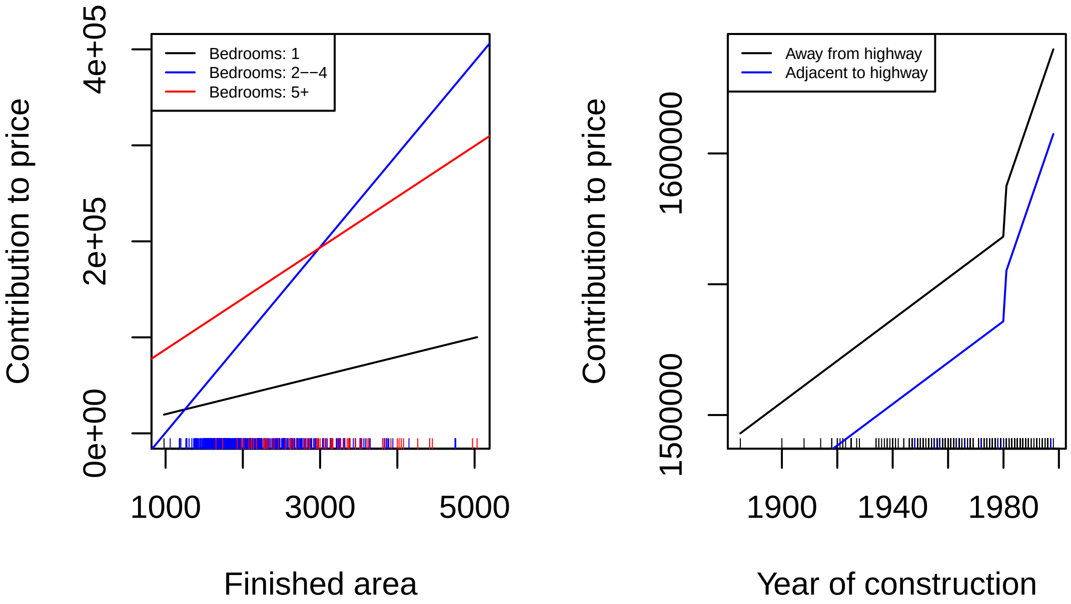

FIGURE 5: *The contribution of finished area to predicted price, as a function of the number of bedrooms (left), and the contribution of year of construction (relative to a baseline of the year 0, which is why the amounts are so huge), reflecting a changing trend after 1980 (right). (The estimate is surprisingly smooth, considering we didn't enforce any sort of continuity.) Rugs along the horizontal axis show the continuous values belonging to each category.*

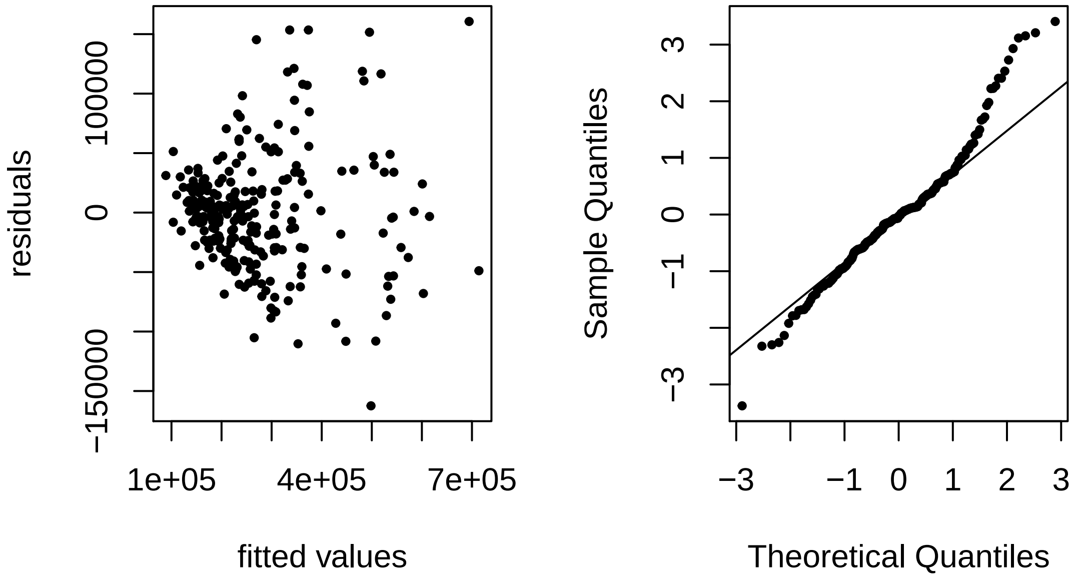

FIGURE 6: *Diagnostics for the final model.*

Afternotes

These are things to think about that weren't required.

Model solution: Remember, this is a model solution. It is not the only solution. You may have included some of these things but not others. The purpose of this *model* is to demonstrate the sort of discussions you should be having and how to articulate them in a reasonable report. Your decisions about what seems "relevant" are likely different than mine, and that's ok.

Data splitting: The procedure I used for model selection here deviates from the "workflow" discussed in class. The reason that I split the data in half and performed model selection on one half was to avoid using the same data to select the model and then produce confidence intervals and p-values. If I had done that, I would have *guaranteed* that the CIs were wrong. In this way, if the selected model is *correct*, the CIs are correct. Whether or not the final model is correct is up for interpretation, but it's not obviously wrong (I don't think).

EDA: With very limited space, it is important to be selective about EDA, and indeed everything else. It is neither necessary nor desirable to include in the report a histogram of every marginal distribution, a description of it, a description of every part of the pairs plot, etc. You should certainly look at those, during your data analysis, but then you should select the subset of plots which will give the reader a reasonable sense of the data, and, more especially, those which actually made a difference to your modeling decisions and analysis.

Initial Model: Here, I deliberately tried to keep the initial model simple, with just what we'd need to answer the client's questions and a few other things which seemed compelling from the EDA. This then needs to be followed by checking whether the variables or terms omitted mightn't have mattered. Another tactic is to throwing everything in initially, and then try to prune the model down. Which one to attempt is largely a matter of taste.

Hunting for Interactions: The number of possible product interactions among p variables is $p(p-1)/2$. This grows far too rapidly for manual examination to be feasible, and even for conventional hypothesis testing. (Throwing in transformations makes things even worse.) You should, therefore, be very selective about hunting for transformations, trying as far as possible to base them on either background knowledge, or things that look funny in the EDA, or the diagnostics on the initial model.

Multiplicative contributions: A log-transformation of the price would lead to a model where each term made a multiplicative, rather than an additive, contribution to the price. Another possibility, a bit trickier to manage with our mathematical tool-kit, would be for price to be proportional to finished area, but at a rate which is a function of the other predictor variables. The model would say, in effect, "high-quality houses built in 1985 near highways sell for so many dollars per square foot" — which sounds very much like a realtor or an appraiser.

Candidate Models: If, when you are looking at your diagnostics and they suggest a change which might improve the model but which is not obviously, overwhelmingly the right thing to do, you have a choice, and each possible choice expands your set of candidate models. Some choices might not work well together, and so you can narrow the pool of candidates by looking at further diagnostic plots. Here, however, I deliberately avoided that, instead just taking all the combinations of the choices (32 of them!) and letting cross-validation sort it out.

Avoiding model selection An alternative to selecting *a* model is to look at many models, rejecting the ones where the assumptions are detectably violated. (You can think of the ones you retain as, almost, a confidence set of models.) You then report a range of predictions, estimates and inferences, from across the retained models. This requires some care (what, *exactly*, are the model assumptions, and *just* how badly can they fit the data before rejection?), and can be harder to explain to clients than picking the best-predicting model.

Grading rubric

Words (4 / 4) The text is laid out cleanly, with clear divisions and transitions between sections and subsections. 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 / 2) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (4 / 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 (5 / 5) 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 / 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 (0 / 0) 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.