Yixin Hu, ECON 21300 Assignment 4

Question 1:

Build an OLS model on your cleaned version of houses_training_set. And then, making whatever adjustments you deem appropriate, provide your best prediction of the sales price for each house in houses_test_1. Describe your thought process and modeling choices. (Answer submissions should be contained in your .csv file with the first column being "parcel_number", the second being "sale_price", and the third being "question_1".)

Response:

First, loading the data:

Getting to know the data, there are a few things that we should note:

It is worth noting that there is duplicate parcel_numbers values. While I believe that a parcel number should be a unique identifier for property, assuming that the fact that there later sale dates for the same parcel number generally has an increasing price and that prices trend upwards over time for real estate, I believe that the most recent price would be must suitable for estimation as it already encapsulates past information. However, since we are expected to estimate each piece of real estate at each date, I will not filter anything out.

From looking at the basic summary of all the datasets, I found that all the school number data have 999s for values. While I do not believe that there is a reasonable explanation for this value, it seems that all school numbers are more explainable using district IDs, where 999 is just a marker of certain schools as well (even if we take the interpretation that 999 is missing data, we can interpret 999 as "those schools not in the identified group")

Specific to the training data, we see that there are 50 counts with -99999s as well as 23 values of 0 within the sales_price data. I don't think that there is a valid explanation besides this was encoding missing data, therefore, I will remove these values in order to make sense of the data as a whole. There is also a quick item to note, which is that there are multiple columns where sale_price is 2600, 24 to be exact. However, other values in the data shows that they do in fact vary with respect to different characteristics, so I do not believe that they are necessarily duplicates. Finally, we also identify that there are 30 NA values in the data, which necessarily makes the data quite literally unpredictable, hence in sum, we will sacrifice these about 100 observations out of 14500, which is a reasonable sacrifice to get the workable data. Not only this, we see from below that the tails are not balanced on both sides. Therefore, for the sake of a better estimation, I will trim the bottom 2% of the testing data with respect to sales price. Furthermore, we see that there are no extreme outliers in the upper end of the distribution of prices, hence we will not remove the top percentiles (see below).

There is also values of -99 in full bathrooms, which I cannot interpret to be anything other than missing values. Therefore, without introducing biases into the data, I will remove these columns (25 in total) to arrive at a reasonable model.

In the training data set, we have one entry where ground floors is 1 sqft, which is flat out ridiculous, as well as 25 values for -999 for second floor area, which does not make sense. Therefore, thus I will remove them accordingly to make a better overall prediction as I cannot interpret this in a reasonable way.

Now looking at street noise, railroad noise, and airport noise. I think it would be helpful to not only look at the training data frame by itself, but also in conjunction with testing data 2, since I would suspect that an increase in street noise would likely mean that the property is situated in a busy urban location, thus usually this would mean higher property prices. However, It seems that for all data frames, the values are either 0 or 60 (61 or street and railroad noise). Looking at the distribution of railroad noise and airport noise, they seem like outliers, and yet looking at street noise, it is more akin to a categorical variable where 0 means no noise, and everything other than 0 means noise. Therefore I will use the street noise as a categorical variable.

At this point, the main data cleaning is tentatively complete, where we actually see that most variables between test 1 and training share similar distributions (which is something that we expect). I will move on to the regression construction and conduct further cleaning should the issues arise.

After cleaning the data, we will begin model construction. To do this, I will further split my training data into a "true training data" and a "validation" set in order to check how my model is doing before estimating the actual testing data no.1.

For the actual model, I believe that there are few variables that are truly significant in determining something as crucial as a real estate's value. In particular, instead of the granular information on how many windows it has, I believe that some of the most important variables lie elsewhere, where living area (i.e. area of floors), maintenance conditions (well-kept vs. run-down), and functionality (bathrooms, garages etc.) are the main "intrinsic" properties which determine the price of a real estate.

Specifically, it is the case that square footage is not the only fundamental determinant of prices, rather,

increased "functionality" (i.e. bedrooms, washrooms etc.) are also extremely important to determine how 'valuable' is a piece of real estate. Lastly, it also should be the case that all else equal, a well-maintained home should be more "valuable" than a run-down home, hence I believe that conditions of the home should also be a factor when considering housing price as well as when the property is constructed. Therefore, I will use area, functionality, and quality to map out the "intrinsic" properties of a piece of real estate. (other bells and whistles such as additional luxuries, pools for example, will be considered later when estimating prices for luxury homes in the second test)

Beyond "intrinsic" properties, and perhaps more importantly, the location of the real estate is extremely important.

It is the case that you can have an extremely luxurious mansion, but if its stuck in a sketchy neighborhood far from where you work, the willingness to pay for the property might not actually be as high as you would have expected. Furthermore while it would be helpful to get data such as local crime rates, mean incomes, and traffic to further estimate the environment/neighborhood of the property, we simply do not have that full set of data, and we are stuck with whatever "external" data that we have. In particular, I believe that schools as a starting point could be helpful in measuring the environment of the neighborhood. This is because access to education often signifies increased school spending in that area, and as research conducted by Barrow et al. (https://www.nber.org/digest/jan03/school-spending-raises-property-values), has a positive effect on per pupil housing values. Furthermore, schools can also serve as a proxy for community stability, as often times amenities (such as basic infrastructure like sewers and clean water, and as well as policing) will be supplied to schools, which should have a positive effect on real estate prices relative to those in less stable communities. (If we also had data on the *quality* of the school rather than the quantity, our estimates would be better as per the Brookings Institute: https://www.brookings.edu/research/housing-costs-zoning-and-access-to-high-scoring-schools/). In sum, I will include schools (high school in particular) as an anchoring point to tease out the neighborhood conditions that the property is located. Related to area, I believe that housing in "special districts" will also be quite important to consider, as it is usually the case that areas with historical significance will positively contribute to sales prices (such as UPC in Hyde Park, where the facilities are not the best, but still cost quite a lot due to it being designed by the same person who designed the Louvre).

Finally, quickly glancing at the real estate website Zillow, we find that among the basic information we have accounted for, the main "Facts and Features" also include heating and chooling. Therefore, this information shows that it is important to also add basic "water and electric type" amenities such as central air, plumbing, and amp ratings to determine sale price as well. There is also the idea of "effective year" vs. "construction year", where "construction year" indicates when the basic structure is built, but does not account for significant renovation or neglect (Florida Real Estate Tax Relief (https://fltaxrelief.com/what-does-effective-year-built-mean-on-the-property-record-pages/)). Therefore I believe having effective year as a true measurement of price would be more ideal as a non-renovated house that is not for sale doesn't necessarily contribute to the price. (i.e. buying a home that has no utilities won't tell us much about homes with utilities)

The first regression will start out simple, I will only consider the main components stated above, i.e. the intrinsic size of the property (areas, functionality, condition) and access to external amenities (traffic, noise, location).

NOTE: FOR ALL QUESTIONS, Please Check code for the full specification, I will only go over particular interesting variables to save space. Nevertheless, all variables are picked with the aformentioned logic in mind.

Here, we have quite a significant and intuitive result when looking at the model at hand. For each square ft increase in assessment area and living area (floor areas), we see a statistically significant increase in housing price. Furthermore, we see that this number makes economic sense as well, as a quick Google search shows that in the Midwest, average cost per square foot is around $106.79. This value is sure to be lower in our analysis because we have added additional covariates to explain real estate prices, and this is reflected in our model. A quick note is that we see additional floors are generally more expensive per square ft than the ground floor, which also makes intuitive sense as an additional sqft for a second story would mean that

you *have* a second story, which will increase prices differently than if that extra sqft is added to a bungalow. The only interesting finding is that dining rooms are slightly significantly different from zero, with a point estimate of negative $7492 and S.E. of $4474. This is quite surprising and does not make any intuitive sense what so ever, as we would expect housing with no kitchens to be worth much less than housing with at least one kitchen. My explanation for this could be the fact that almost every single entry has at least 1 kitchen, and there are only 32 out of 13663 entries in the cleaned training data that does not have a kitchen value of 1. Therefore, it could very likely be the case that the "unpredictiveness, or unrepresentativeness" of the kitchen variable is due to the fact that practically every single entry has at exactly the same value. Lastly, it should be noted that price is also positively increasing with respect to effective year, which makes sense since effective year is when it is actually livable, more recent the renovation means higher the price.

Now looking at the 'external' factors combined with square footage, I have selected the access to schools for this analysis. By looking at availability of schools, we see that high schools have a positive and statistically significant result, where high school raises prices by a (quite large) value in the millions, *except* for one district (149) that actually has a negative effect on price (and district 142 with insignificant results). Glancing at the data, we see that there are not a lot of housing in 149 and it could be the case that the housing around school 149 is not necessarily the best property for selling, and those that are sold tend to sell for less. Therefore, this hints at the neighborhood conditions around school 149. Furthermore, as expected, we see that property in a national historic district has a positive effect on price by a factor of $32670. In the end, this model has an R^2 of around 0.7794.

Now onto a predicting sample with the "validation" set

In looking through our validations, there are a few odd pieces of data that is quite interesting. First of all, a regression between the true price and the predicted price has a coefficient point estimate on the prediction of 0.989, which means that our predictions itself does not overestimating or underestimating the prices systematically. Particular for our validation set, after we have tuned the model, we have obtained a MSE of 480118633, and the distribution of the estimated prices are also given below as a sanity check (see if there are extreme values in our predictions). To make sure there is no significant overfitting, our true training set has an MSE of 475995806. Furthermore, running a regression of our estimate on the RHS and the true price on the LHS, we see that our point estimate is 1.014 with a S.E. of 0.01. All previous indications shows that there is no significant overfitting in our model. Finally, we move on to the predictions of the first testing set.

Question 2

Build a RANDOM FOREST model on your cleaned version of houses_training_set. And then, making whatever adjustments you see fit, provide your best prediction of the sales price for each house in houses_test_1. Describe your thought process and modeling choices.

Response:

Building the random forest model, we will use whatever insight that was obtained previously to proceed with the model. Before we start, it is worthwhile to note that random forest models inherently has variable selection, which to some extent removes the requirement for my own judgment. However, I think it would not be idea if we just let the model do whatever it wants. To address this fully, I will experiment with different model specifications and provide the most sensible model as my final result. In doing so, I am able to (at least) get some sense as to what is going on behind the scenes and comment on the way.

Tuning the model a few times, I have noticed that the original set of predictors used in OLS actually approximates the validation set very well. This is to be expected of a RF model. However, what I lose in this situation is the intuition behind this model. In particular, I do not know why or how does the RF model actually selects the variables and assigns what types of weights on each variable. Therefore, besides the fact that we are running a consistent model based on my discussion outlined in question 1 of variable and modeling selection, there is not much more to talk about in terms of truly understanding the model.

After tuning the model, we have arrived at a good MSE value for RF at mtry = 7. Now, we can make a clear comparison between the MSE of OLS estimates and MSE of RF estimates from the same true testing set and the validation set. In particular, recall that we have MSE of OLS = 482056869, whereas the MSE of RF = 382577164. This shows that the RF performs better than the OLS when testing against our validation set. It is of course important to note that these values will differ from different sampling of true training sets and validation sets. Furthermore, using our predictions to explain true prices, we see that we retrieve a point estimate of 1.078 and an S.E. of 0.009. Which means that our model does not see any systematic underestimating or overestimating. Now, to make the predictions for the first set of testing data using the RF model

QUESTION 3:

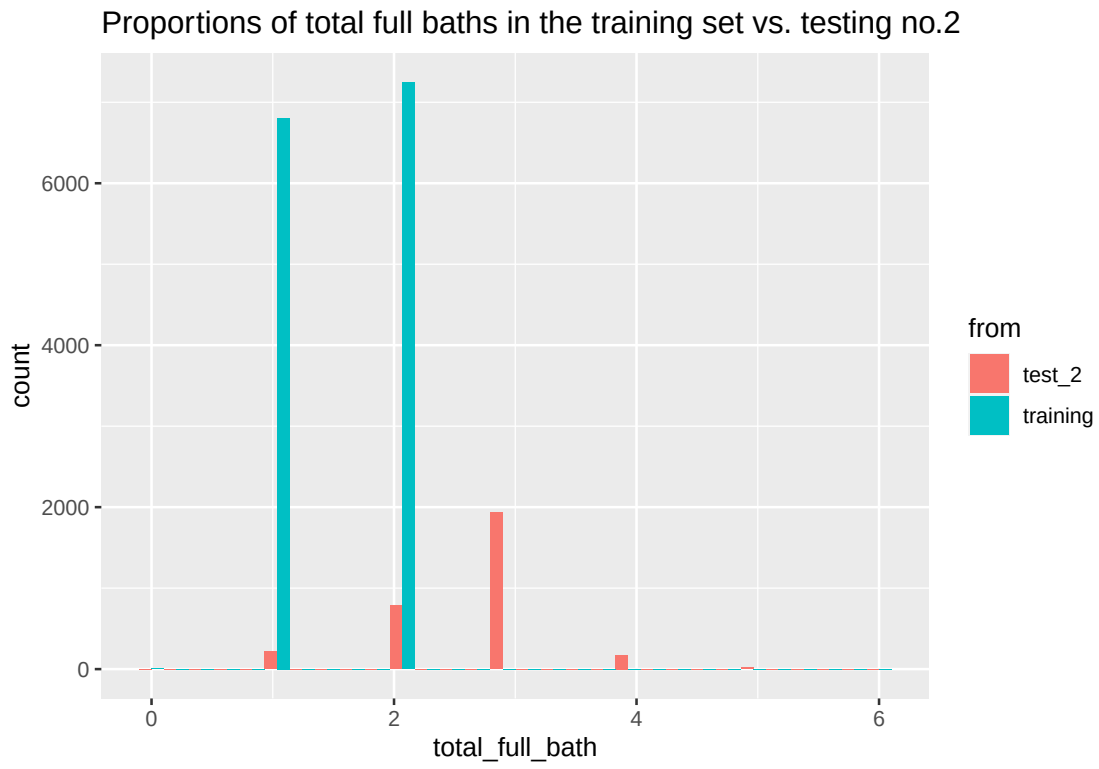
Using whatever modeling approach and whatever adjustments you deem appropriate, provide your best prediction of the sales price in houses_test_2. Describe your thought process and modeling choices. NOTE: you have to be thoughtful here!!

Response:

There are two things that we have to consider when doing this part.

- $300,000+ Sale price homes are now included
- 2 or fewer full bathrooms are now included

Therefore, we must somehow take into consideration how these two factors will affect our estimations. First checking the distribution of full bathrooms vs. the training set...



There are two peices of information to get from this graph: 1) the training data does have much more rows of data relative to the testing no.2, and 2) Indeed, we see that most of the full bathroom in the second testing set is centered around 3 bathrooms, which we cannot estimate directly from the training set. However, recall that in our OLS method, we did not use this data as a factor (fixed effect), and treated it linearly. This makes sense because it is common in the real estate circle to use the amount of full bathrooms to estimate how luxurious a property is (if a smaller bathroom will do this trick, having a full bathroom is simply for luxury), therefore, the OLS method seems to have a better chance at predicting correctly relative to RF methods as we are saying "how does an **additional full bathroom** increase price", and thus have a theoretical meaning behind our estimates.

Exploring the data of our indicated set of variables supports our "luxury" assumption. By checking the key groups of factors: intrinsic and external, we see that distribution of area of property for all floors in the 2nd testing set is shifted to the right and has a thicker right tail, indicating an increase in larger homes. They

have more garages, increased quality class, higher AMP ratings. However, at the same time test set 2 has similar inside and outside conditions, similar central air, and similar sale dates. However, I think that to expand and correctly estimate luxury homes, we should include more variables considering "non-essential" characteristics, such as fireplace openings to squeeze out some predictive power in estimating luxury homes.

At the end, it is as if we are blind to the "luxurious data" and we are estimating their property using only information that we see in our normal, non-luxurious data. This intuitively means that what might be considered "valuable" for us, such as good plumbing, might be completely an after thought in higher-end housing since it is already become a "given" for them. It is **not** the case that they pay "less" for the basic amenities, but it should be the case that the *bulk of their value comes from somewhere else* that we do not observe clearly in our "non-luxurious" data. Therefore, if we are to only look at our training data, it is almost for sure that we will be **underestimating** their housing prices, since it is as if "good pluming" is the limit of determining value and completely oblivious to the value given by, say, a backyard tennis court. In this light, it must be the case that we will be *biased* in our estimation, where we will give systematically lower estimations than what will be true. I think a good way to account for this shift is to make a "shift" parameter which adjusts our biases in the predictions **OVERALL**.

Now moving into particularities about full bathrooms, notice that we have cases where housing with 2 or less bathrooms are included in the data, which means that these must be priced \$300,000 or *HIGHER*. On the other hand, we could also possibly have cases where properties have high full bathrooms, but possibly priced *LOWER* than \$300,000. This would mean that our full bathroom point estimate as a blanket statement is quite unreliable. i.e., While it is true that in our previous testing that full bathroom is a great predictor of price, our new data is definitely going to skew the point estimate on the full bathroom variable. In the case where we have introduced housing that has \$300,000 price and yet lower full bathroom counts, this would mean that our predictor of full bathroom would increase, whereas the case where we have introduced housing that has less than \$300,000 but higher full bathrooms, the predictor of full bathroom would decrease.

While one possible way to address this problem is to lump the two variables together. However, since I have checked that the two variables are independent (by checking their interaction terms), lumping the two variables into a joint "bathroom" effect would not be a sound procedure as we are basically saying two full bathrooms have the same effect on price as two half bathrooms.

Therefore, the best thing that we can do model wise is to work with our OLS assumptions that even in the situations where we have the out-of-bounds full bathroom data. As will be outlined later, we are making the assumption that, in accordance with linear estimates of OLS, **our growth ratio is expected to be linear in nature, a.k.a the distributions increase at a roughly constant geometric value**, i.e., the ratios of growth from 1 full washroom to 2 full washrooms in terms of 1 full washroom is the same as growth from 2 full washrooms to 3 full washrooms in terms of 2 full washrooms. Therefore, what we can do is to check the growth of the distributions of 1 full washroom to 2 full washrooms in our training data and assume that all subsequent number of full bathrooms grows in the same geometric fashion (constant ratios).

Indeed, we can take training data so that we can make a "mock test" where we split our training data into a validation set above \$200,000, and a training set of under \$200,000 in sales price. This is to mimic the arbitrary cut off and check how does the *bias* introduced in sampling.

To do this, I will split the training data into two segments: "high-cost" housing and "normal" housing. This is to replicate the situation we have here with test set 2. Then I will list a bunch of assumptions to make our logic consistent with our model. After constructing our model, I will go ahead and create the shift parameter which measures our inherent bias in predictions, and use this constructed parameter as a hyper parameter (i.e. something we choose ourselves) to tune our predictions accordingly.

Here, we have taken an arbitrary cut off to mimic the situation where we must use the full training data to approximate the testing set 2. Doing so allows us to investigate a few aspects of our model which will be useful in approximating. To recap, we are told that our data is *biased* in terms of pricing, as well as the variable on full bathrooms are *out-of-bounds* in the testing data relative to our training data. Therefore, the aspects that we are trying to estimate in our "mock test" is:

- to what degree are the estimates biased in terms of sales price on a whole, and thus coming up with a

so called “shift parameter”

- to what extent do increased full bathrooms change the distribution in sale prices, and thus coming up with a “distribution multipliers” for each extra full bathroom

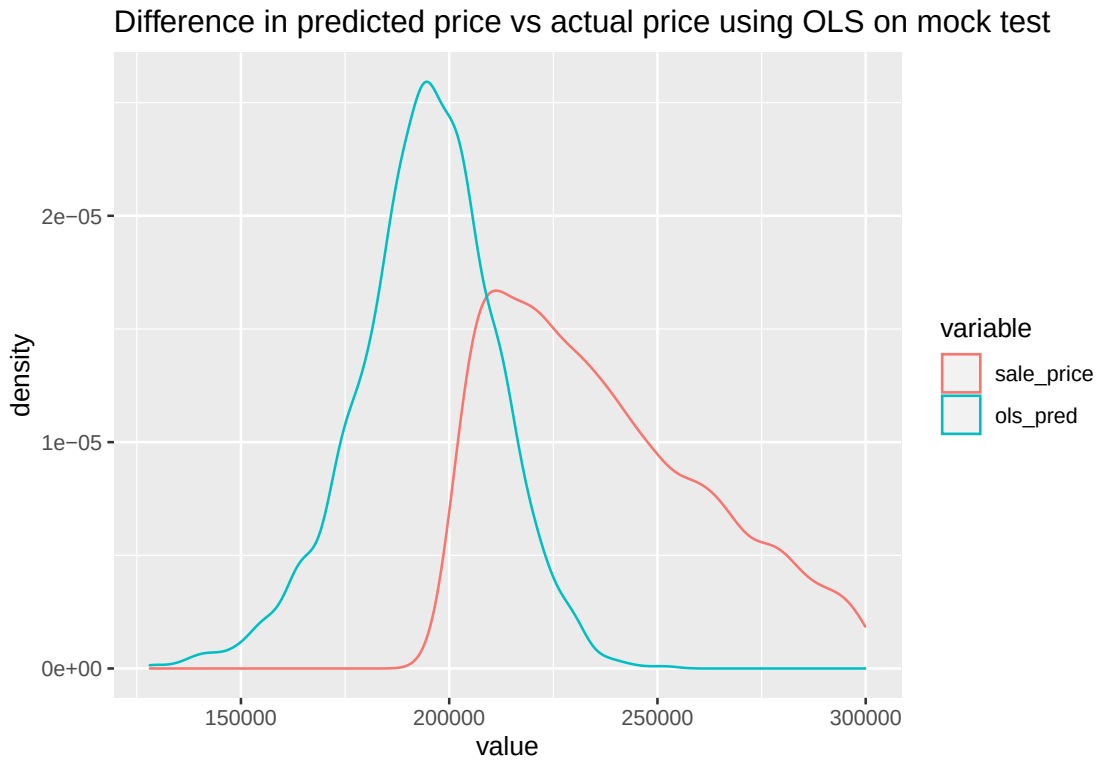
Needless to say, we have some key assumptions when using these values. In tuning the model, we are making the following assumptions:

- the training data set’s properties are “less-luxurious” than the second testing set’s data in terms of prices and full bathrooms
- between the two data sets, point estimates on fundamental characteristics are still relatively the same (i.e. going from 500 sqft to 501 sqft increases the value of any property by a similar amount, all else equal, between the two sets of data)
- **the bias caused by predicting “more-luxurious” housing and “less-luxurious” housing is relatively consistent across groups of property with similar characteristics (shift parameter)**
- **additional functionality provided by additional bathrooms contribute to a change in prices by a relatively consistent ratio in terms of distributions’ mean and standard deviations (distribution multipliers)**

First to find the sale_price distribution differences between different amounts of full baths. (Note, 0 is left out for this graphic as there are only 5 rows with 0 full baths)

Here we see from both the graphic and the table that increasing full baths will shift the distribution of sale prices to the right, and increase its standard deviation by a slight amount. To be precise, we find that the two means are separated by 1.25x the one-full-bath housing, and the standard deviation varies by 1.06x. Assuming that all subsequent distributions of increasing full baths will have a similar growth factor, I will take these values to tune my predictions once the model is complete.

Now to building the model on our mock test to check the bias in the model.



From the summary on the true price regressed on our prediction, we see that we are getting a far lower value than our previous estimates, specifically, we have that the point estimate is 0.621 and an S.E. of 0.02. Furthermore, we have a more direct visualization where I plot the true sales price and the predicted price, and the distributions are completely off. What is happening here is that we are taking a *biased* sample to estimate something out-of-bounds. Therefore, we need to account for the *biasedness*.

After running a few different arbitrary cut-offs around 200,000, we have confirmed that the shift parameter should be around 1.22. This means that we were almost always 1.22 around times underestimating the biased sample with OLS (with lowest values of 1.19 and highest values of 1.27 in some of my tuning). Therefore, to account for this bias, it would then be the case that we are making another linear assumption: we are assuming that in our arbitrary cutoff of 300,000, we will by systematically underestimating by approximately 1.22 times as well, as per our "mock test". I fully acknowledge that this is extremely empirical and arbitrary, yet I simply cannot think of any other theoretically sound way to measure how an out-of-bounds test will do besides running a mock test on what data we already have.

To sum up, running some tests in our "mock test", we find that our OLS estimates using the training data below the arbitrary cut offs is systematically underestimating the prices, where the true prices are on average 1.22x times higher than our predicted results for multiple cutoffs, meaning that our shift parameter should be around 1.2x. Looking at the distributions of prices between 1 full bathroom vs 2 full bathrooms, we find that the mean of 2 full bathrooms is 1.25x higher than the mean for 1 full bathroom, and the standard deviation is 1.07x higher in terms of standard deviation, meaning that these two values will provide the distribution multipliers 1.25 for the mean and 1.07 for the standard deviation.

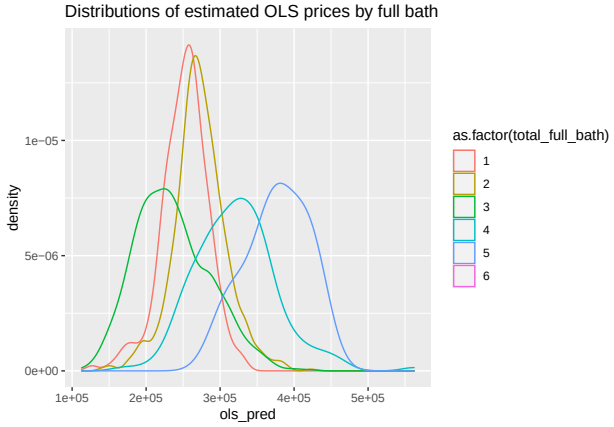
However, it is also worth noting that even after all this tuning, if we find cases where full bathrooms < 2

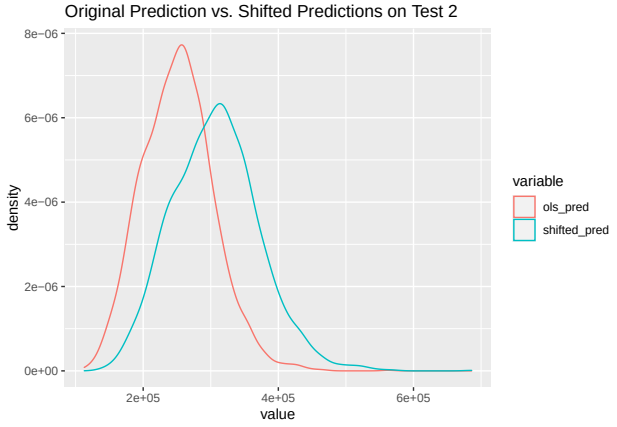
AND sale price $<$ /\$300,000, it would mean that we would of course be incorrect since this data would have never even appeared in our testing set 2. Therefore, I will set these values to \$300,000 as it is the lowest (and thus closest value to our estimation) value which satisfies the conditions on our test set 2.

Therefore, the algorithm for finding our best predicted price is the following:

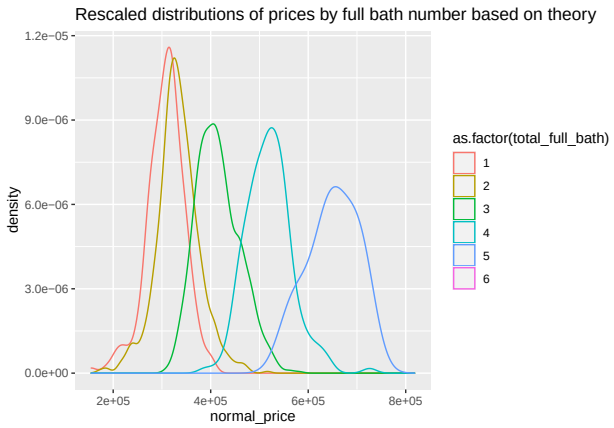
- First estimate the entire testing set 2 using OLS.
- Then, I will consider the overall bias and shift the predicted prices by 1.2x.
- Then, I will have to consider the distributions of prices for each amount of full bathrooms to see if anything is needed in terms of the distribution, and apply our distribution multipliers accordingly.
- Finally, I will consider the items where it would have never existed in the testing set 2 and rescale their estimations to /\$300,000 to comply with the data description

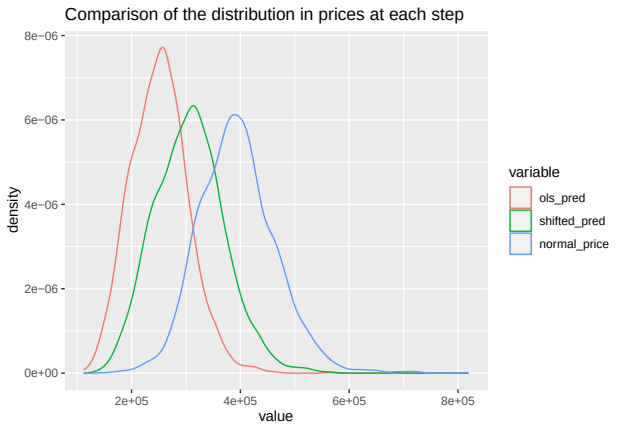
Now that we have our assumptions and algorithm, we can begin to create our model for test set 2 using our full training data.





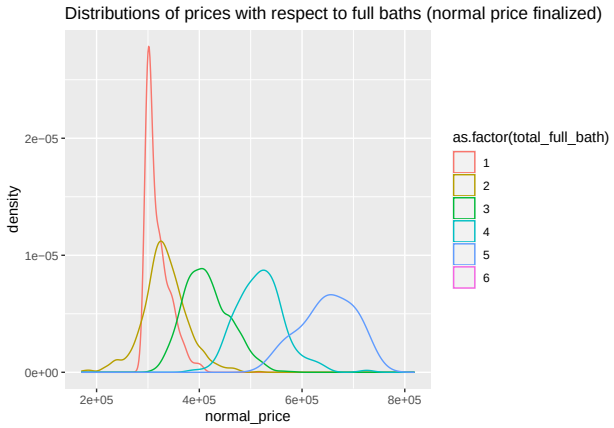
Notice that our resealed estimates looks much more beleiveable, as we have attempted to remove the bias from our inherently biased sample. However, the predictions when looking at full bathrooms are completely counter to our theory that each successive full bath will increase the mean of the price distribution by 1.25x and the standard deviation by 1.06x from 2 washroom onward. Hence, we will scale the values accordingly, without touching the values below 2. In the graphics below I present the rescaled distributions together to get a sense of how exactly is the distribution shifting, as well as a graphic to show how applying relevant multipliers and shifts to the distributions of the prices with respect to full baths based on our empirical observations, assumptions will get us a distribution that is far more aligned with our theory. In particular, 'normal_price' is the price when we applied multipliers with respect to each full bath levels, where OLS predictions and shifted predictions are simply the predictions constructed before.

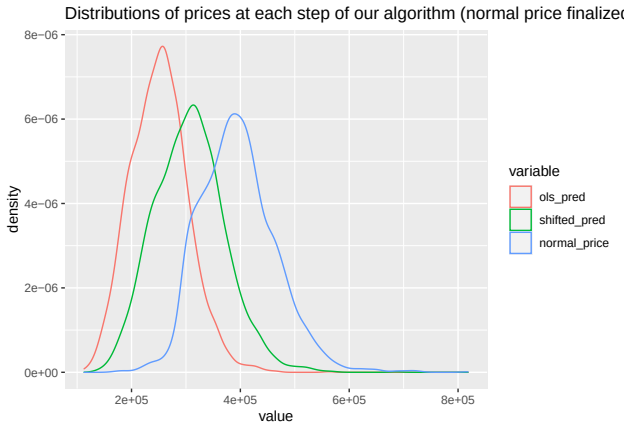




Notice then our shifted values are more aligned with our theory, where each successive increase in full baths will increase the mean and standard deviations accordingly. Now to consider the case where we are estimating housing that is below $300,000 and yet still only has less than or equal to 2 bathrooms. To take care of this, the only reasonable thing to do is to scale them all to exactly $300,000. This is because this is the closest value that is indicated within our data description, and we can only make such a guess in this context. Notice that the normal_price now will reflect this change, and thus give us our finalized prediction.

NOTE: I fully acknowledge the very real dangers of overfitting here. In fact, I would even argue that we are definitely overfitting to some extent. However, I simply cannot think of any other way to take into account the sheer uncertainty that necessarily surrounds out-of-bounds testing besides taking what we already know about the data, and applying it to our predictions. While some may argue that this is completely invalid in the sense that we have used overfitted arguments, I would push back and say that at least we are using the characteristics of what we have learned from the true data in order to make our best educated guess as to how unobservables will affect price.





Now that we have made all necessary adjustments, the final estimates (normal_price) will be provided in the csv and renamed to sale_price with question 3 attached.

The final exported csv will be “yixinhu_assignment4_predictions.csv”