KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics Project

Project Name: Predicting Housing Prices

Class Section: ITEC-621-001 (Thursday)
Team Number: Team 2
Team Members:
Peter Chika Ozo-Ogueji
Michael Guddah
Loc Le
Nga Thi Quynh Ho

Last updated: 30 April 2024
Deliverable Number: 4

1. Business Question and Case

***1.1. Business Question*:** For real estate agents or individuals looking to sell a house in Seattle, Washington, what are the most important home features to consider when predicting the sales price?

***1.2. Business Case*:** According to experts, the housing market is currently in a unique position as despite higher mortgage rates, home prices are rapidly rising at record levels due to a lack of housing supply and increased demand (Ostrowski, 2024). In fact, U.S. home prices in February 2024 increased by 6.5% in comparison to last year (Redfin, 2024) while the overall housing supply remains low when compared to historical standards (Carbonaro, 2024). As a result, sellers have a lucrative opportunity as experts predict that there will be an increase of homes selling for more than the asking price (Martin, 2024). Given the unfavorable market conditions for buyers, those looking to purchase a home will undoubtedly be seeking to maximize their value. Therefore, by analyzing the most important features of a house, sellers can learn what aspects significantly contribute to its value. This information can then be leveraged to develop a robust predictive model for predicting the sales prices of a home which is beneficial as it allows sellers to understand their position in the market and therefore set competitive listing prices. This not only ensures optimal profitability but also resonates with buyers seeking maximum value in their investment, thus attracting more overall buyers. With this approach, a balance can be struck between seller profitability and buyer satisfaction which will ultimately drive success in the fiercely competitive housing market.

2. Analytics Question

2.1. Analytics Question: What effect does the most important features of a house have on its selling price? By uncovering the most significant features that influence the selling price of a house (e.g., size, layout, condition, etc.), sellers can set competitive pricing strategies that maximize profitability while maintaining value for buyers. Therefore, our analytics goal will be focused on both predictive accuracy and interpretation as we aim to maximize the performance and reliability of our models while also being able to clearly communicate our findings to real estate agents and individuals looking to sell a house.

2.2. Outcome Variable of Interest: The outcome variable of interest will be price (USD) which is quantitative, continuous, and represents the selling price of an individual home.

2.3. Main Predictors: The main predictors for developing the predictive model will likely include important house aspects such as land and property size, number of rooms and floors, listing condition, and age.

3. Data Set Description

The Seattle housing price dataset, sourced from Kaggle, encompasses 21,613 entries with 21 variables. Each variable offers distinct insights into housing characteristics and pricing which facilitates a comprehensive understanding of Seattle's real estate market. Key variables such as “price” (numeric), “bedrooms” (numeric), “bathrooms” (numeric), “sqft_living” (numeric), “sqft_lot” (numeric), and “floors” (numeric) provide fundamental details about property size and value, while features including “waterfront” (binary) and “view” (categorical) offer insights into amenities and scenic qualities. Additionally, variables such as “condition” (categorical) and “grade” (categorical) offer assessments of property quality, while “yr_built” (categorical) and “yr_renovated” (categorical) shed light on historical and renovation aspects (see Appendix A1 for the key variable definition dictionary).

4. Descriptive Analytics

4.1. Descriptive Statistics of Key Variables: Our outcome variable for this Seattle housing price analysis is “price” (USD) which has a median of $450,000 and a mean of $540,084. This suggests that the distribution is likely skewed to the right as that is generally expected when the mean is larger than the median. Therefore, the median likely provides a more robust measure of central tendency as the mean is more susceptible to influence from higher-priced outliers. The price variable also has a standard deviation of $367,135 which suggests that there is relatively high variability as many data points are

located further from the mean. As for our key predictor variables, "bedrooms", which represents the number of bedrooms in a house, has a median of 3, a mean of 3.369, and a standard deviation of 0.93. "Bathrooms", which represents the number of bathrooms in a house, has a median of 2.25, a mean of 2.115, and a standard deviation of 0.77. "Sqft_living", which represents the total square footage of a house, has a median of 1910, a mean of 2080, and a standard deviation of 918.44.

4.2. Distribution of Key Variables: Our initial exploration exposes a key characteristic of the price variable which is that it does not follow a normal distribution and is right-skewed instead. This is indicated in the histogram and the QQ plot where the data points deviate from the QQ line towards both ends of the tail, especially towards the right end. Given the results from these visualizations, it appears that the dataset contains several homes that were sold for very high prices. While these may initially appear as outliers, that can only be confirmed with further analyses as it is normal for there to be significantly less expensive homes (e.g., mansions) than regular homes in a city such as Seattle. As for our numeric key predictor variables, upon examining their histograms and QQ plots, the same right-skewed trend is present in all of them (see Appendix B1 for several histograms and QQ plots).

4.3. Correlation and Covariation Analysis: To understand how various features influence price, we investigated correlation coefficients. The correlation matrix reveals the relationships between the house price and other quantitative variables of interest (see Appendix B2 for the correlation matrix). "Sqft_living" exhibits the highest positive correlation with price (0.70), indicating that larger living areas tend to command higher prices. While not as strong, "bathrooms" (0.53), also has a moderate positive correlation with price. Conversely, "house_age" (-0.05) has the lowest correlation with price, suggesting a weaker association with the age of a house. As for the categorical variables of interest, our ANOVA tests indicate that price varies across different groups for "waterfront", "view", "condition", and "grade" as they produced significant p-values that were greater than 0.05 (see Appendix B3 for boxplots).

4.4. Data Pre-Processing and Transformations: In the data pre-processing and transformations phase, our first step involved checking for data cleanliness where we initially looked for missing values but didn't find any. We did, however, identify a house that had 33 bedrooms, and upon further investigation, we deemed this specific data entry to be an unnatural outlier which resulted in its removal from the dataset. Afterward, we decided to create the "house_age" variable (how old a house is), which was calculated by subtracting the year each property was built ("yr_built" variable) from the current year, as we believed that property age has an impact on price where older houses are expected to have lower prices. However, this expectation generally only applies if the house has not been renovated. Therefore, to reflect the renovation status of each home, we examined the year a house was renovated ("yr_renovated" variable) and created a new binary variable called "renovated" where 0 indicates no renovation ("yr_renovated" = 0) and 1 otherwise ("yr_renovated" = any year). Once we created our two variables, we removed the variables used to make them as well as "id", "date", "sqft_above", "sqft_basement", "zipcode", "lat", "long", "sqft_living15" and "sqft_lot15" as we deemed those to be unnecessary for the analysis. Lastly, we converted all the remaining variables to their proper data types where "waterfront", "view", "condition", "grade", and "renovated" were converted to factors as they are categorical, while the numerical variables were already in their proper formats and did not require any conversions (see Appendix A2 for the data pre-processing steps and an overview of the dataset).

5. Modeling Methods and Model Specifications

5.1. Initial Model Specification: In order to begin predicting house prices in Seattle, we created an initial OLS model consisting of "price" as our quantitative outcome variable and because our objective is to find the most important features for predicting sales prices, we decided to include all the variables deemed relevant to the analysis as our predictors. This includes numerical predictors that represent the number of bedrooms ("bedrooms"), bathrooms ("bathrooms"), and floors ("floors"), a house's age ("house_age"), as well as the total square footage of the entire property ("sqft_lot") and the area that is liveable ("sqft_living") which are all important physical house specifications. Also included are categorical predictors that represent important assessment ratings such as a house's overall scenic

outlook (“view”), condition (“condition”), and appraisal score (“grade”). In addition to those, there are also binary predictors that indicate whether or not a house has a waterfront (“waterfront”) and if it has been renovated (“renovated”). While this initial OLS model contains many predictors, we expect that some of them will be removed as we check for assumptions and make any necessary corrections, as well as perform variable selections, best subsets, and other methods.

5.2. Initial OLS Results: Upon fitting our initial OLS model using the 11 chosen predictors, we immediately noticed that all of the numerical predictors were significant with p-values less than 0.05 as well as all the levels (minus the baseline) for the categorical predictors of “waterfront”, “view”, and “renovated”. On the other hand, while nearly all levels for “condition” and “grade” have p-values greater than 0.05 and are therefore not significant, the higher levels for each variable (condition5, grade11, grade12, and grade13) are significant with p-values that are less than 0.05. Furthermore, this initial OLS model explains around 68% of the variability in sales prices for houses sold in Seattle as demonstrated by the R-squared value (see Appendix C1 for full summary results).

5.3. Assumption Tests: The first OLS assumption test is met as it requires that the response variable, “price”, is continuous (EY). Conversely, the second OLS assumption test regarding the normal distribution of errors (EN) is violated as there is a deviation from the QQ line at both ends of the tail. Based on the correlation matrix examined previously, we already had suspicions of multicollinearity due to correlation values among several predictors being greater than 0.5. To confirm this, we checked the condition index and received a staggering CI of 581.91 which normally suggests that the VIF for each predictor should be examined. However, due to the OLS model containing dummy variables with multiple levels, the GVIF is supposed to be examined instead and upon checking those values for our model, no predictors exceeded a GVIF value of 5. Therefore, the OLS assumption test regarding independent predictors (XI) is met. The OLS assumption test regarding linearity (LI) is also met as all predictors appear to be linearly related to “price” based on individual scatterplot analyses that indicate no substantial departures from the straight lines. Because our analysis does not involve time series, the OLS assumption tests for independent observations and errors (OI and EI) are not required for our analysis. Additionally, the 0 error average OLS assumption test (EA) is not necessary as our model has an intercept. Lastly, the OLS assumption test regarding homoscedasticity (EV) appears to also be violated as there appears to be an uneven spread of the residuals (see Appendix B4 for visualizations of OLS assumptions).

5.4. Model Candidates and Rationale: Our first set of models utilized OLS and WLS modeling methods. Due to the two violated assumptions (EN and EV) in the initial OLS model, we decided to build a subsequent log-linear model where the response variable (“price”) was logged and was indeed able to correct for the non-normal distribution of errors. After constructing the log-linear model, we then used it to try and correct the heteroskedasticity issue by fitting a weighted least squares (WLS) model. However, upon checking to see if the heteroskedasticity issue was resolved with the WLS model, we found that not only was the issue still present, but the errors were also no longer normally distributed. As a result, we tried additional weighting reiterations to see if the violations could be corrected but the issues seemed to get worse after each iteration. Therefore, we decided to pivot to the LASSO regression method for our second set of models as LASSO models are less sensitive to the violated OLS assumptions we couldn’t correct. We also chose LASSO over Ridge regression as LASSO is capable of shrinking coefficients to 0 which aligns with our goal of finding the most influential features for the selling price of a house. Despite LASSO being able to better handle the violations, we decided to use Random Forests for our final set of models as it is a non-parametric method where the OLS assumptions and other parametric restrictions no longer apply.

5.5. Model Specification Candidates and Rationale: In addition to logging the “price” response variable for each model method candidate (OLS/WLS, LASSO, and Random Forests), we also used two different sets of predictors for each modeling method as our other specifications. This resulted in “full” models consisting of the initial 11 predictors that were selected from a business knowledge perspective and “small” models which only consisted of 9 predictors (“bedrooms”, “bathrooms”, “sqft_living”, “floors”, “waterfront”, “view”, “condition”, “grade”, and “house_age”). These 9 predictors were chosen using

stepwise variable selection (both directions) where a null model was used for the lower bound and a full model was used for the upper bound. Additionally, a p-value of 0.05 was set as the threshold to only include the most significant predictors. Our rationale for doing this is because we wanted to see if any of the 11 predictors we initially selected were not important for influencing the sales price of a house and whether or not the models would improve in terms of predictive accuracy if the two unimportant variables ("sqft_lot" and "renovated") were removed. Without considering the initial OLS model where "price" is not logged, we ultimately constructed 10 models where there were two that used OLS (OLS.full and OLS.small), two that used WLS (WLS.full and WLS.small), two that used LASSO without weights (LASSO.full and LASSO.small), two that used LASSO with weights (LASSO.full.wts and LASSO.small.wts), and two that used Random Forests (RF.full and RF.small) (see Appendix D1 for a table and diagram that clearly details all the modeling methods and specifications used).

5.6. Cross-Validation Testing: In order to evaluate the performance of all 10 predictive models, 10-fold cross-validation was used as the testing method where the mean squared errors (MSE) were calculated for each model as the statistical comparison metric. For the OLS models, the full specification (OLS.full) and small specification (OLS.small) yielded MSE values of 0.097169 and 0.097129, respectively. However, when these models had weights added to them, there were minor changes in the MSE values as the full specification's (WLS.full) MSE increased to 0.097179 while the small specification's (WLS.small) MSE also increased to 0.097135. As for the LASSO models, we obtained MSE values of 0.097110 for the full specification (LASSO.full), 0.097091 for the small specification (LASSO.small), 0.097089 for the weighted full specification (LASSO.full.wts), and 0.097070 for the weighted small specification (LASSO.small.wts). Lastly, while there have only been miniscule differences in the MSE values between the parametric models thus far, the non-parametric Random Forest models yield much better results as the full specification (RF.full) and small specification (RF.small) have MSE values of 0.086300 and 0.091526, respectively (see Appendix D2 for a table that contains the MSE and RMSE values for all models).

5.7. Final Model Selections: In regards to our analytics goal of both predictive accuracy and interpretation, we have decided to select two different models for each goal. For predictive accuracy, we selected the full random forests model (RF.full) as it had the lowest MSE value of 0.086300 which is much smaller than all other models. However, despite having the best predictive performance, this model could not be selected for our other goal as random forest models are generally not good for interpretation. Therefore, we decided to select the small OLS model (OLS.small) for interpretability as even though it has a larger MSE than the LASSO models, it is still acceptable due to the differences only being minute and OLS models are generally the best for interpretation.

6. Analysis of Results

Starting with the full random forest model (RF.full) which is the best for predictive accuracy, this model explains approximately 70% of the variance in house prices as demonstrated by the R-squared value and according to the variable importance analysis, "house_age", "sqft_lot", and "grade" were the most influential predictors in predicting house prices, while "renovated" had the least impact. As for the small OLS model which we deemed to be the best model for interpretation, it explains around 65% of the variance in house prices and has an acceptable 10-fold cross-validation test MSE value of 0.097129. Furthermore, it is worth mentioning that while the normal distribution of errors assumption is met in this model, the constant error variance assumption is still violated (see Appendix C2 for the full results of the two final models after being fit on the entire dataset). And to answer the analytics question, the following interpretations have been made based on the coefficients from the small OLS model to uncover the effects that the most important features have on a house's selling price:

Quantitative predictors (all interpretations are "on average and holding all other variables

constant"): Among the quantitative predictors, the number of bathrooms is one of the effects that are significant on house prices where each additional bathroom results in an 8.4% increase in house price. Additionally, the total square footage of a home that is liveable and the number of floors also significantly affects house prices where it increases by 0.018% and 7.7%, respectively, for each additional square foot of living space and additional floors. A house's age also has a significant positive effect on house prices, with a 0.6% price increase for each additional year. Lastly, one interesting finding that might seem counterintuitive at first glance is that, for each additional bedroom, there is actually a 3.3% decrease in house price. This effect, however, is still significant.

Binary predictors (all interpretations are "on average and holding all other variables constant"): Because the renovation status of a house has been removed in the smaller model, "waterfront" is the only binary predictor present. It also has a significant effect where houses with waterfronts lead to a 31.5% increase in house price in comparison to those without waterfronts.

Categorical predictors (all interpretations are "on average and holding all other variables constant"): Starting with a house's overall scenic outlook which was scored on a scale of 0-4 where 0 was set as the reference level, the effect of each remaining level was found to be statistically significant and positive where house prices would increase by 18%, 9.7%, 12.7%, and 25.2% if the house's view was scored to be 1, 2, 3, and 4, respectively, compared to those with a score of 0. As for a house's overall condition, the effect of houses who received a score of 3, 4, and 5 are positive and significant, with price increases of 14.9%, 16.8%, and 23.4%, respectively, in comparison to houses with condition scores of 1. The effect for houses with condition scores of 2, however, is negative and not significant where those prices are expected to decrease 1.4% when compared to houses with condition scores of 1. Lastly, the effects for the overall grade a house receives, which is scored between 1-14 (no houses received a score of 2), are relatively mixed where approximately half of the levels are significant while the rest are not. For the effects that are not significant, when compared to houses with grades of 1, houses with grades of 3, 4, and 5 decrease the price by 5.1%, 5.7%, and 1.1%, respectively, while houses with grades of 6 and 7 increase the price by 20.8% and 48.7%, respectively. For the effects that are significant, houses with grades of 8, 9, 10, 11, 12, and 13 increase the price by 72.2%, 95.6%, 111.5%, 123.1%, 131.9%, and 134%, respectively, when compared to houses with grades of 1

7. Conclusions and Lessons Learned

7.1. Conclusions from the Analysis: To address the business question, we shifted our focus back to the full random forest model (RF.full) and the variable importance plot where the most important features of a home when predicting its sales price, such as a house's age, the total square footage of the property, and the appraisal grade it receives, were generally what we expected based on our understanding of the real estate market. While not as influential as those three, the total liveable square footage, how good the view is, the number of floors, bathrooms, and bedrooms, and overall condition were still all at variable importance levels that made sense. The two least important aspects, however, were surprising to us as we expected houses with waterfronts and new renovations to be desirable features that would normally lead to higher prices. Nevertheless, when it comes to predicting a home's sales price in Seattle, Washington, we can conclude that the most important features to consider are how old a house is, how big the entire property is, and the appraisal grade it receives. With that being said, our full random forest model can serve as a useful tool for real estate agents or individual home sellers in providing accurate sales price predictions for houses based on their specific features.

7.2. ***Project Issues, Challenges and Lessons Learned:*** The main challenge we faced in this project had to do with addressing the violated assumptions in the initial OLS model as we could not successfully correct them both. Although this was not difficult to address as we were told early on that we should progress to other modeling methods where OLS assumptions are not as sensitive, we were persistent in trying to find a solution that would simultaneously correct both. As a result, we ended up spending a lot more time with this issue than we should have. Additionally, we also ran into several coding errors in RStudio where some were easily resolved while others were more difficult. For instance, the issue that took us the longest to resolve was when we were trying to do 10-fold cross validation on the WLS models using the coding examples we learned from class. For some reason, we received multiple errors during this process despite following the code and relevant online resources and when one error would be resolved, another one would occur. Once we got to a point where we could not resolve the issue without restructuring the entire dataset which would drastically change all of our results and force us to start back at the beginning, we found alternative 10-fold cross validation coding methods online and were immediately were able to use them to produce results without any errors. One important lesson learned from this experience is connected with the first challenge we discussed and is how OLS models can still provide value even if the assumptions are violated. While we still understand the importance of testing and meeting these assumptions, violations are not catastrophic issues that absolutely require corrective action and in most cases, the results can still be useful as long as the violations are accounted for in the insights.

Appendices

A. Data Information

A1.

Definition dictionary of key variables:

- Price - The price at which a house is sold
- Bedrooms - The number of bedrooms a house has
- Bathrooms - The number of bathrooms a house has
- Sqft_living - The total square footage of the property that is liveable (i.e., the house)
- Sqft_lot - The total square footage of the entire property including the house, lawn, backyard, driveway, etc.
- Floors - The number of floors in a house
- Waterfront - A binary variable that indicates whether or not a house has a waterfront view (0 = no waterfront and 1 = waterfront)
- View - The score given to a house based on its scenic outlook (ranges 0-4)
- Condition - The score given to a house based on its overall condition (ranges 1-5)
- Grade - The grade given to a house after an appraisal (ranges 1-14)
- Yr_built - The year a house was built
- Yr_renovated - The year a house was renovated
- House_age - The age of a house in years
- Renovated - A binary variable that indicates whether or not a house has ever received renovations (0 = no renovations and 1 = has been renovated)

A2.

R code used for data preprocessing:

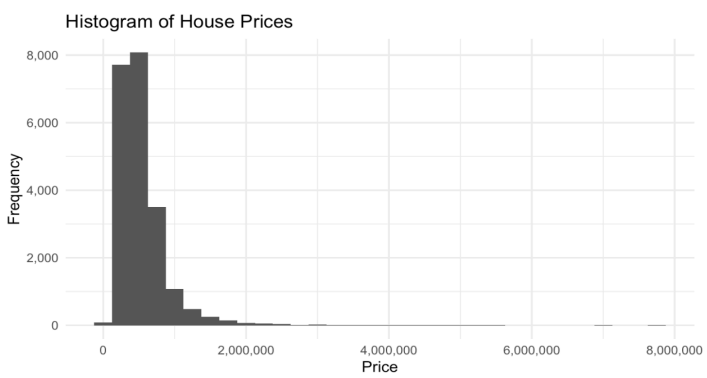
```{r}
# Read in dataset
seattle_housing <- read.csv("house_sales.csv")

# Check for NA's
na_count <- sum(is.na(seattle_housing))
print(paste("Number of missing variables:", na_count))

# Create house_age variable
date <- as.numeric(format(Sys.Date(), "%Y"))  # Get current year
seattle_housing$house_age <- date - seattle_housing$yr_built

# Create "renovated" variable
seattle_housing$renovated <- ifelse(seattle_housing$yr_renovated == 0, 0, 1)

# Remove unneccesary variables by creating data subset
seattle_clean <- seattle_housing %>%
  select(price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront,
view, condition, grade, house_age, renovated)

# Data type conversions
seattle_clean$waterfront <- as.factor(seattle_clean$waterfront)

seattle_clean$view <- as.factor(seattle_clean$view)

seattle_clean$condition <- as.factor(seattle_clean$condition)

seattle_clean$grade <- as.factor(seattle_clean$grade)

seattle_clean$renovated <- as.factor(seattle_clean$renovated)

# Remove outlier house with 33 bedrooms
seattle_clean <- seattle_clean[-15871, ]
```

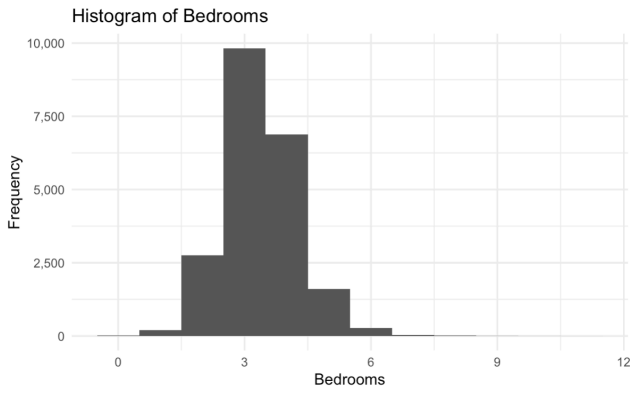
Brief overview of the fully cleaned dataset using the glimpse() function:

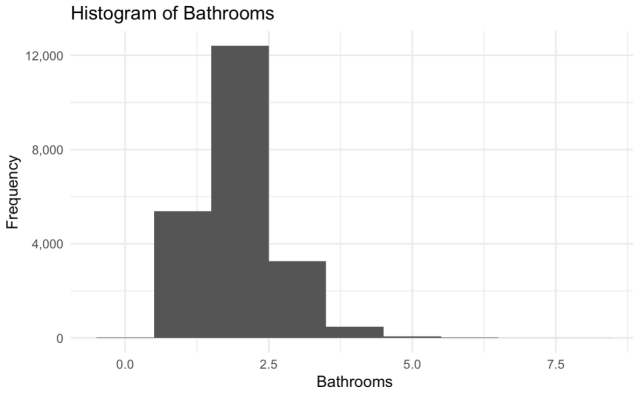
```
Rows: 21,612
Columns: 12
$ price       <int> 221900, 538000, 180000, 604000, 510000, 1225000, 257500, …
$ bedrooms    <int> 3, 3, 2, 4, 3, 4, 3, 3, 3, 3, 3, 2, 3, 3, 5, 4, 3, 4, 2, …
$ bathrooms   <dbl> 1.00, 2.25, 1.00, 3.00, 2.00, 4.50, 2.25, 1.50, 1.00, 2.5…
$ sqft_living <int> 1180, 2570, 770, 1960, 1680, 5420, 1715, 1060, 1780, 1890…
$ sqft_lot    <int> 5650, 7242, 10000, 5000, 8080, 101930, 6819, 9711, 7470, …
$ floors      <dbl> 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1.…
$ waterfront  <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ view        <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, …
$ condition   <fct> 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 4, 4, …
$ grade       <fct> 7, 7, 6, 7, 8, 11, 7, 7, 7, 7, 8, 7, 7, 7, 7, 9, 7, 7, 7,…
$ house_age   <dbl> 69, 73, 91, 59, 37, 23, 29, 61, 64, 21, 59, 82, 97, 47, 1…
$ renovated   <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
```

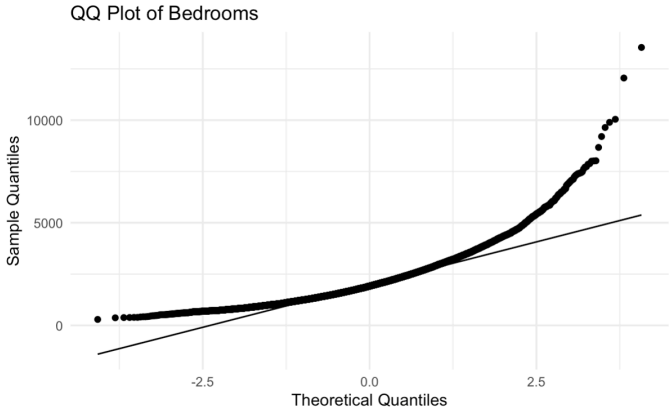
B. Visuals, Graphs and Plots

B1. Histograms and QQ plots of several key variables:



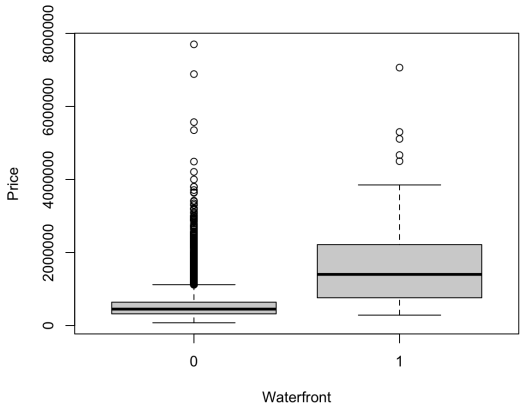


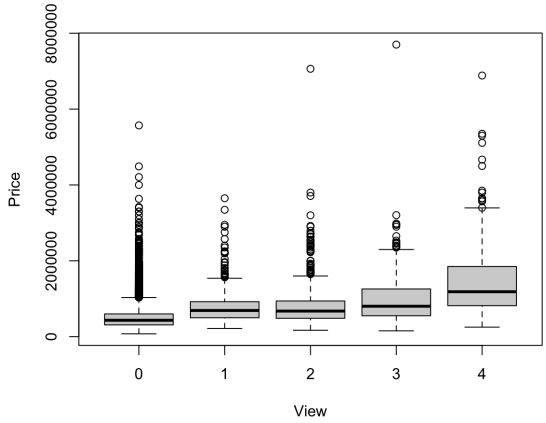


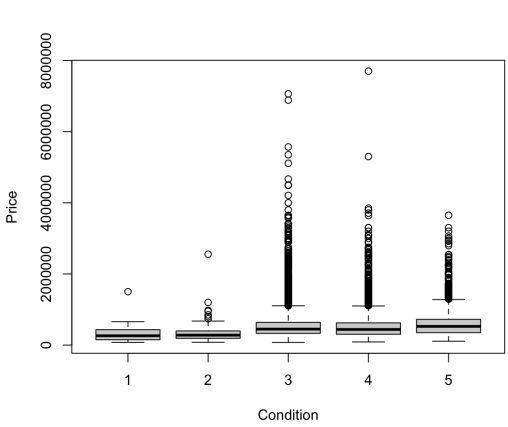


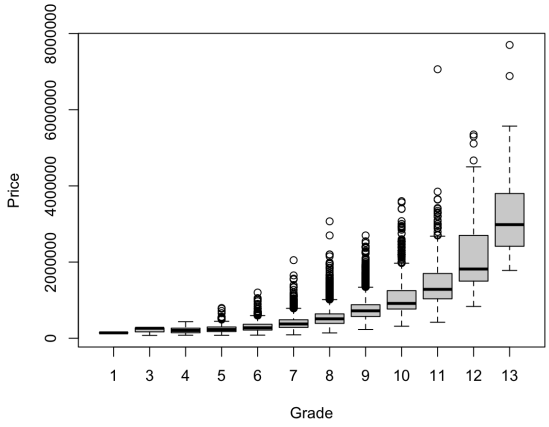
B2. Correlation matrix visualization of quantitative variables:

B3. ANOVA Boxplots for categorical variables:



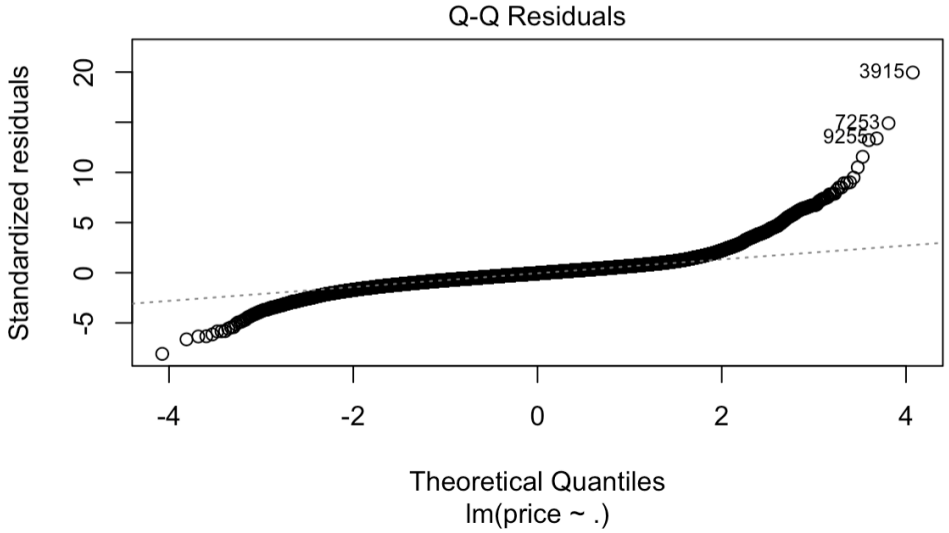






B4. Visualizations for initial OLS model assumptions:

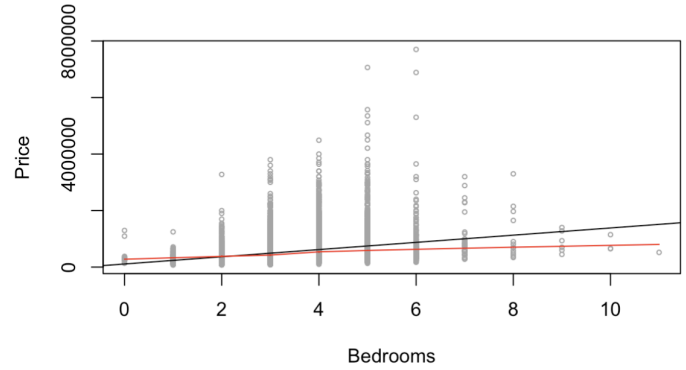
OLS Assumption (EN) - Errors are normally distributed:

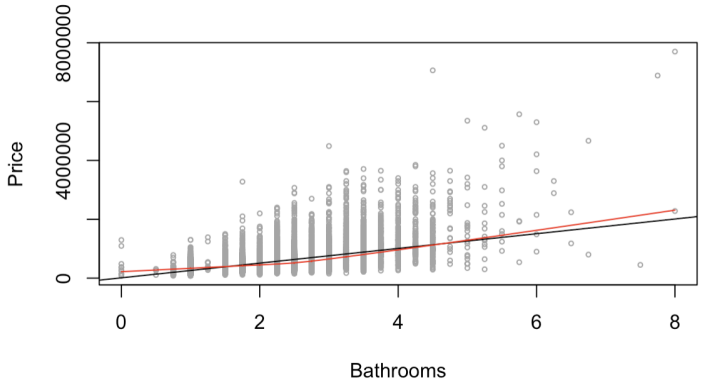


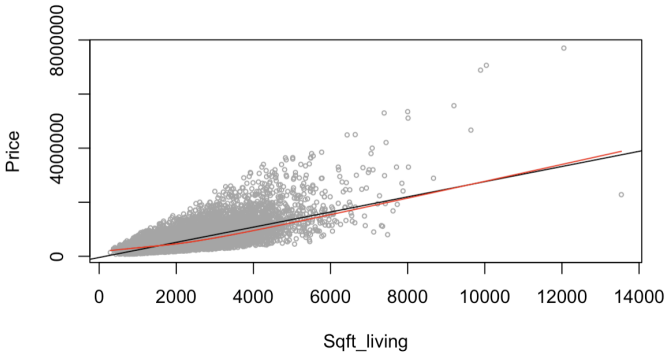
OLS Assumption (XI) - Independent Predictors:

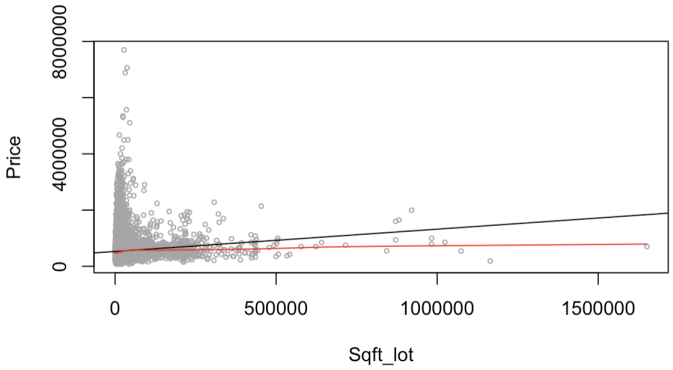
	GVIF	Df	GVIF^(1/(2*Df))
bedrooms	1.777740	1	1.333319
bathrooms	3.314017	1	1.820444
sqft_living	4.587856	1	2.141928
sqft_lot	1.056761	1	1.027989
floors	1.628214	1	1.276015
waterfront	1.551707	1	1.245675
view	1.765777	4	1.073660
condition	1.389062	4	1.041934
grade	3.839419	11	1.063059
house_age	2.071222	1	1.439174
renovated	1.154983	1	1.074701

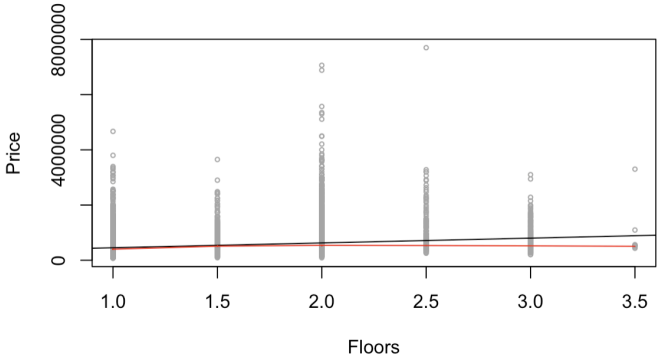
OLS Assumption (LI) - Linearity:

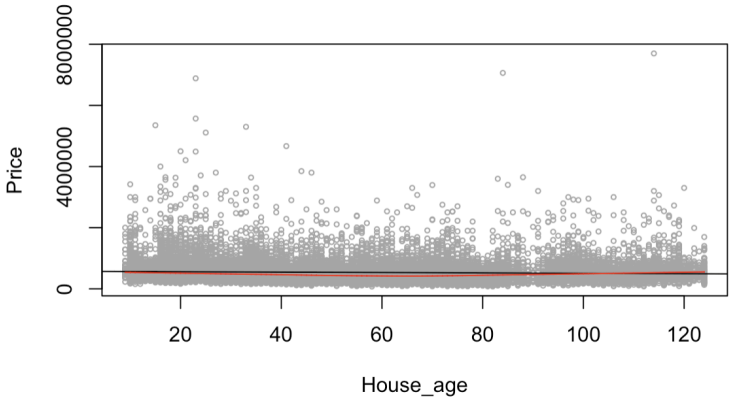




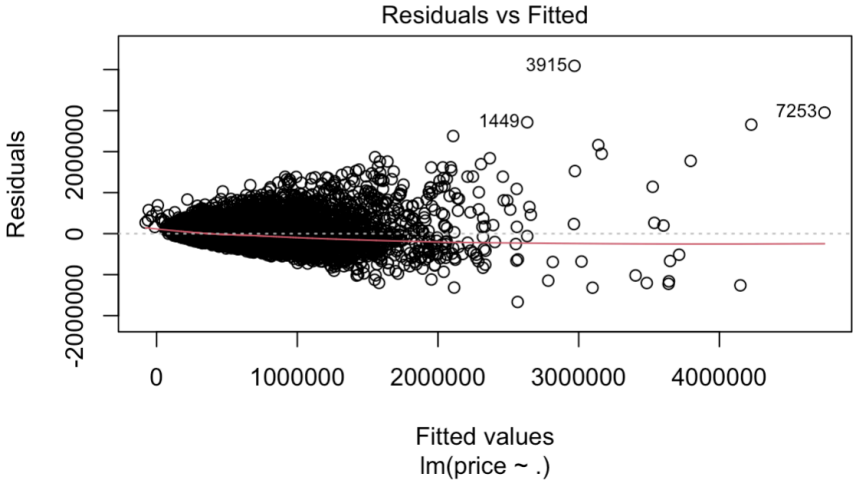








OLS Assumption (EV) - Homoscedasticity:



C. Quantitative R Output

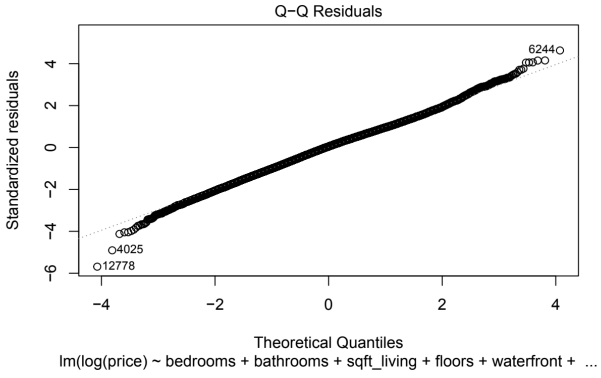
C1. Initial OLS model coefficient results:

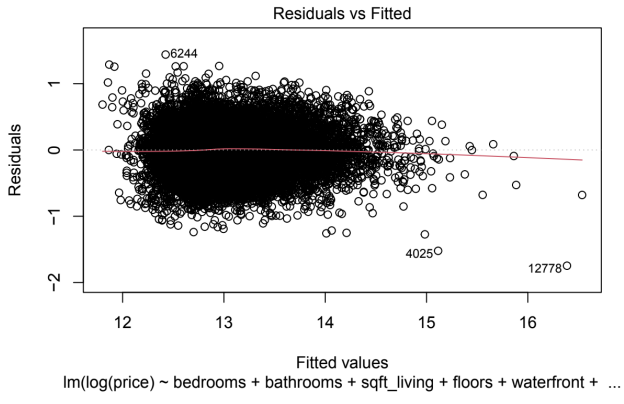
```
Call:
lm(formula = price ~ ., data = seattle_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1666192  -106071   -11105    86038  4092466

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.196e+05  2.074e+05  -0.577  0.56412
bedrooms    -2.673e+04  2.071e+03 -12.907  < 2e-16 ***
bathrooms    5.115e+04  3.333e+03  15.345  < 2e-16 ***
sqft_living  1.375e+02  3.289e+00  41.802  < 2e-16 ***
sqft_lot    -2.662e-01  3.500e-02  -7.607 2.92e-14 ***
floors       3.174e+04  3.332e+03   9.524  < 2e-16 ***
waterfront1  4.888e+05  2.030e+04  24.076  < 2e-16 ***
view1        1.228e+05  1.157e+04  10.612  < 2e-16 ***
view2        5.669e+04  7.003e+03   8.095 6.02e-16 ***
view3        1.097e+05  9.572e+03  11.461  < 2e-16 ***
view4        2.595e+05  1.480e+04  17.534  < 2e-16 ***
condition2   1.395e+04  4.175e+04   0.334  0.73832
condition3   3.902e+04  3.887e+04   1.004  0.31546
condition4   5.761e+04  3.887e+04   1.482  0.13834
condition5   9.794e+04  3.909e+04   2.506  0.01223 *
grade3      -6.987e+04  2.426e+05  -0.288  0.77330
grade4      -1.138e+05  2.142e+05  -0.531  0.59518
grade5      -1.390e+05  2.111e+05  -0.658  0.51046
grade6      -8.761e+04  2.110e+05  -0.415  0.67796
grade7      -5.454e+03  2.110e+05  -0.026  0.97938
grade8       8.312e+04  2.110e+05   0.394  0.69369
grade9       2.249e+05  2.111e+05   1.066  0.28666
grade10      4.063e+05  2.112e+05   1.924  0.05441 .
grade11      6.720e+05  2.115e+05   3.178  0.00148 **
grade12      1.128e+06  2.125e+05   5.309 1.11e-07 ***
grade13      2.323e+06  2.194e+05  10.589  < 2e-16 ***
house_age    3.206e+03  6.910e+01  46.400  < 2e-16 ***
renovated1   3.339e+04  7.531e+03   4.435 9.27e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 207300 on 21584 degrees of freedom
Multiple R-squared:  0.6815,    Adjusted R-squared:  0.6811
F-statistic:  1711 on 27 and 21584 DF,  p-value: < 2.2e-16
```

C2. Final Model Results

OLS.small:

```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + sqft_living +
    floors + waterfront + view + condition + grade + house_age,
    data = seattle_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74661 -0.20576  0.01512  0.20882  1.44002

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.138e+01  3.111e-01  36.574  < 2e-16 ***
bedrooms    -3.295e-02  3.096e-03 -10.643  < 2e-16 ***
bathrooms    8.043e-02  4.957e-03  16.225  < 2e-16 ***
sqft_living  1.845e-04  4.882e-06  37.784  < 2e-16 ***
floors       7.733e-02  4.986e-03  15.508  < 2e-16 ***
waterfront1  3.145e-01  3.042e-02  10.336  < 2e-16 ***
view1        1.803e-01  1.735e-02  10.388  < 2e-16 ***
view2        9.658e-02  1.051e-02   9.193  < 2e-16 ***
view3        1.270e-01  1.435e-02   8.855  < 2e-16 ***
view4        2.518e-01  2.220e-02  11.343  < 2e-16 ***
condition2  -1.427e-02  6.264e-02  -0.228 0.819748
condition3   1.489e-01  5.830e-02   2.555 0.010630 *
condition4   1.677e-01  5.831e-02   2.875 0.004045 **
condition5   2.342e-01  5.864e-02   3.994 6.52e-05 ***
grade3      -5.117e-02  3.639e-01  -0.141 0.888175
grade4      -5.690e-02  3.214e-01  -0.177 0.859462
grade5      -1.061e-02  3.168e-01  -0.033 0.973280
grade6       2.076e-01  3.165e-01   0.656 0.511883
grade7       4.870e-01  3.166e-01   1.538 0.123955
grade8       7.227e-01  3.166e-01   2.283 0.022466 *
grade9       9.557e-01  3.167e-01   3.018 0.002551 **
grade10      1.115e+00  3.169e-01   3.519 0.000435 ***
grade11      1.231e+00  3.172e-01   3.882 0.000104 ***
grade12      1.318e+00  3.188e-01   4.136 3.55e-05 ***
grade13      1.340e+00  3.292e-01   4.070 4.71e-05 ***
house_age    5.785e-03  9.764e-05  59.246  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.311 on 21586 degrees of freedom
Multiple R-squared:  0.6517, Adjusted R-squared:  0.6513
F-statistic:  1615 on 25 and 21586 DF,  p-value: < 2.2e-16
```

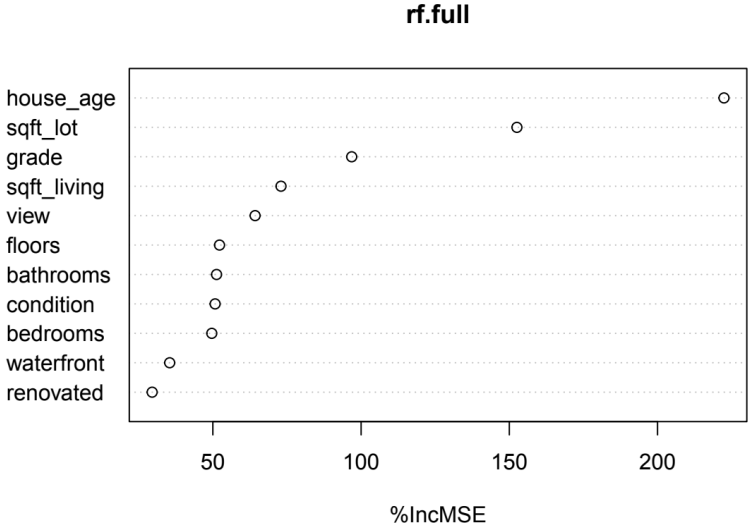




Errors are normally distributed (EN) but heteroskedasticity (EV) is still present.

RF.full:

```
Call:
 randomForest(formula = log(price) ~ ., data = seattle_clean,      mtry = 4, importance = T)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 0.08438146
                    % Var explained: 69.58
```

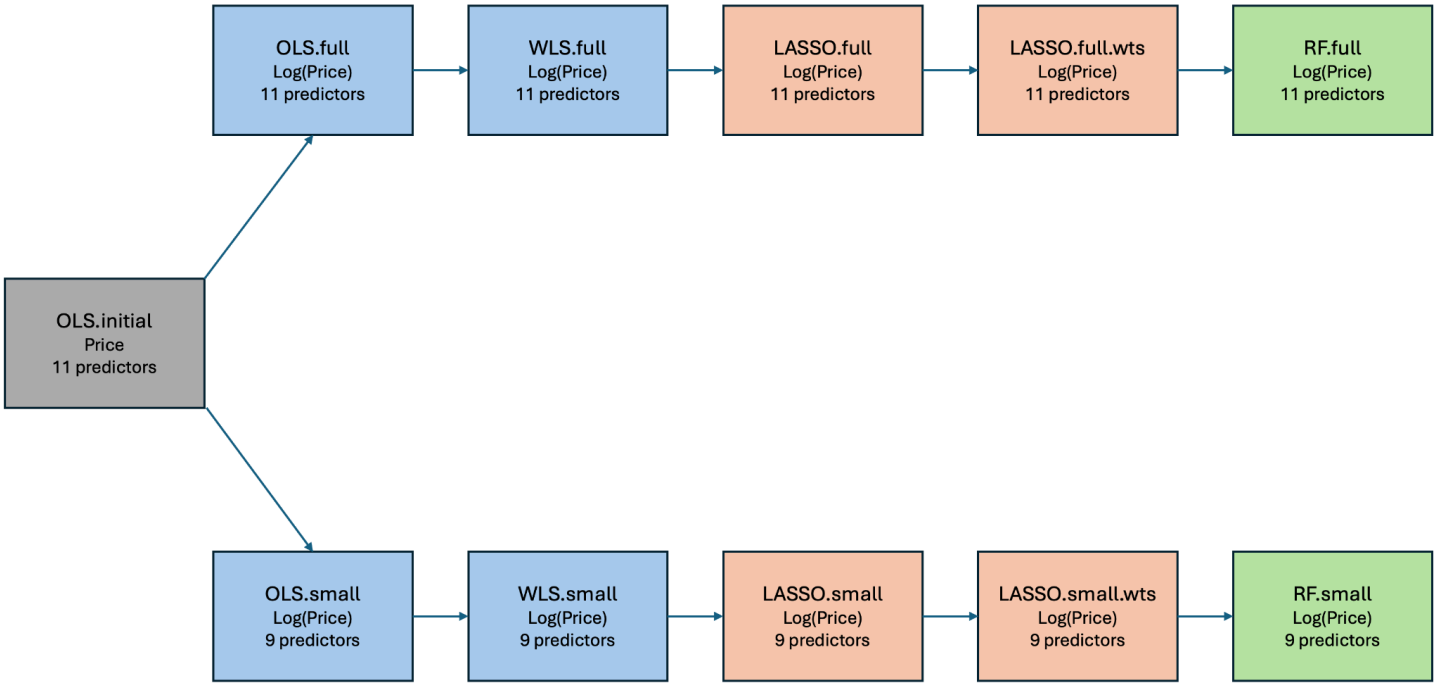


	%IncMSE
bedrooms	49.63291
bathrooms	51.24166
sqft_living	72.98404
sqft_lot	152.55696
floors	52.24112
waterfront	35.42999
view	64.24916
condition	50.75915
grade	96.82785
house_age	222.40691
renovated	29.51846

D. Other

D1. Table and diagram of the modeling methods and specifications used

	Model Names	Methods	Specifications	Response Variable	Number of Predictors	Assumptions/Tests	Notes
Modeling Method 0	OLS.initial	OLS	- Business knowledge	Price	11	EN Assumption (❌) EV Assumption (❌)	- Parent model - Does not count towards requirements
Modeling Method 1	OLS.full	OLS	- Business knowledge - Log(Y)	Log(Price)	11	EN Assumption (✅) EV Assumption (❌)	- Main purpose is to correct EN assumption and use for stepwise
	OLS.small	OLS	- Stepwise - Log(Y)	Log(Price)	9	EN Assumption (✅) EV Assumption (❌)	- Main purpose is to find variables for subsequent small models
	WLS.full	OLS/WLS	- Business knowledge - Log(Y)	Log(Price)	11	EN Assumption (❌) EV Assumption (❌)	- Weight calculated from OLS.full - Reiteration does not fix violations
	WLS.small	OLS/WLS	- Stepwise - Log(Y)	Log(Price)	9	EN Assumption (❌) EV Assumption (❌)	- Weight calculated from OLS.small - Reiteration does not fix violations
Modeling Method 2	LASSO.full	LASSO	- Business knowledge - Log(Y)	Log(Price)	11	No Multicollinearity (✅)	- Unweighted
	LASSO.full.wts	LASSO	- Business knowledge - Log(Y)	Log(Price)	11	No Multicollinearity (✅)	- Weighted - Weight calculated from OLS.full
	LASSO.small	LASSO	- Stepwise - Log(Y)	Log(Price)	9	No Multicollinearity (✅)	- Unweighted
	LASSO.small.wts	LASSO	- Stepwise - Log(Y)	Log(Price)	9	No Multicollinearity (✅)	- Weighted - Weight calculated from OLS.small
Modeling Method 3	RF.full	Random Forest	- Business knowledge - Log(Y)	Log(Price)	11	No Assumptions	- No weight
	RF.small	Random Forest	- Stepwise - Log(Y)	Log(Price)	9	No Assumptions	- No weight



D2. Table of the MSE and RMSE 10FCV results for each model

Model	MSE	RMSE
OLS.full	0.097169	0.311720
OLS.small	0.097129	0.311655
WLS.full	0.097179	0.311736
WLS.small	0.097135	0.311665
LASSO.full	0.097110	0.311625
LASSO.full.wts	0.097089	0.311591
LASSO.small	0.097091	0.311595
LASSO.small.wts	0.097070	0.311560
RF.full	0.086300	0.293768
RF.small	0.091526	0.302532

E. References

Carbonaro, G. (2024). Blue States Are Creating a Housing Market Crisis. *Newsweek*. https://www.newsweek.com/blue-states-housing-market-crisis-1877226

Mahajan, S. (2019). Seattle House Sales Prices. *Kaggle*. https://www.kaggle.com/datasets/sameersmahajan/seattle-house-sales-prices

Martin, E. J. (2024). 2024 Second-quarter Housing Trends: High Hopes for Spring. *Bankrate*. https://www.bankrate.com/real-estate/housing-trends/

Ostrowski, J. (2024). Is the Housing Market Going to Crash? What the Experts Are Saying. *Bankrate*. https://www.bankrate.com/real-estate/is-the-housing-market-about-to-crash/

United States Housing Market. (2024). *Redfin*. https://www.redfin.com/us-housing-market