Predicting Housing Prices Based on Other Characteristics

Anmol Lakhotia, Kyle Bistrain, and Saanvi Dua

California Polytechnic State University, San Luis Obispo

Winter 2023

Table of Contents

II.	Introduction	5
III.	Materials and Methods	6
IV.	Split the Data	7
V.	Data Visualization	7-9
VI.	Variable Pre-Processing	10-13
VII.	Residual Analysis	10-13
VIII.	Fit a Linear Model	14-15
IX.	Statistical Inference	16-18
X.	Model Validation	19
XI.	Conclusion	20
XII.	Appendix	21
XIII.	Reference Page	22

Acknowledgments

We would like to thank our professor, Dr. Bret Holladay, for providing us with the learning opportunity that this project brings, as well as the knowledge and support to complete it successfully. Additionally, we are grateful for an enlightening and engaging quarter, despite the constraints brought on by the weather and the pandemic. We also hope you make lovely unforgettable memories with your daughter and wish you extra hours of sleep.

Group Photo

II. Introduction

Hedonic regression is an important topic in financial research that uses statistical methods to determine a property's value by analyzing how its features affect the value of the property as a whole. It is frequently used in the real estate industry to explain how various characteristics, like location, size, and amenities, impact the price of a home.

In 1974, economist Sherwin Rosen conducted a study on the value of housing in Chicago. By using hedonic regression to analyze features of houses, such as the number of bedrooms, proximity to schools, and access to public transportation, Rosen was able to estimate the value of each feature and how it contributed to the overall price of the house [1]. In our project, we aim to do a similar task, however in response to houses sold in the state of Washington during the summer of 2014. According to Zillow in 2021, the size of the housing market in the United States is estimated to be around $36.2 trillion [2]. Research in housing is important as the housing market is a major driver of the US economy, and its size and performance have significant impacts on the overall health of the economy. Both the common house buyers and sellers benefit from this research as it allows for informed pricing decisions for both to be made.

III. Materials and Methods

We acquired a comprehensive housing dataset from Kaggle.com that offers a wide variety of variables that enables us to conduct an accurate and appropriate regression analysis. The information from the dataset allows us to construct a basic regression model capable of generating reliable predictions.

In our study, the houses serve as observational units, with the primary variable of interest being the price (in dollars). Additionally, we are interested in several other variables, including the number of bedrooms, living space area (in square feet), number of floors, waterfront (0 for no, 1 for yes), and the year it was built. The response variable we are investigating is the price of the house. Although we were unable to find information on how the data was recorded, we assume that it was most likely obtained through random sampling. It is worth noting that all of the housing data included in our study is from the State of Washington, and we do not have any further details regarding the study design.

IV. Split the Data

The housing dataset from Kaggle had 4600 observations. However, when understanding the value of the variables in context, we realized we had to do filtration. In the dataset, there were houses that sold for under $8,000, but in reality, a house would not sell for that much. We also decided to filter houses that sold for over $1,000,000 due to the fact that the value from the property was likely more for the land that it was on than the house itself, i.e. farmland. Although *sqft_lot* would help explain this, the focus of this paper and regression is on housing, not a property that includes ranches, farming, and commercial land. The last criterion that we filtered upon was the year that the house was built. This was due to the fact that older houses are built with different safety features and regulations than newer houses. In the United States, the National Housing Act of 1934 established the Federal Housing Administration (FHA) and set new standards for home construction and lending practices [3]. Due to this, we decided to filter out all houses before the year 1934. After the filtration process, our housing dataset was reduced to 3634 observations.

After that, our training dataset contains 2908 observations, and our test dataset contains 726 observations. We used 111 for our set.seed() function. It is important to split the data so that you can test your model made with the trained data on a test dataset that we have never seen before. This allows us to verify whether or not our model is valid on a random subset of your population data in an unbiased manner.

V. Data Visualization

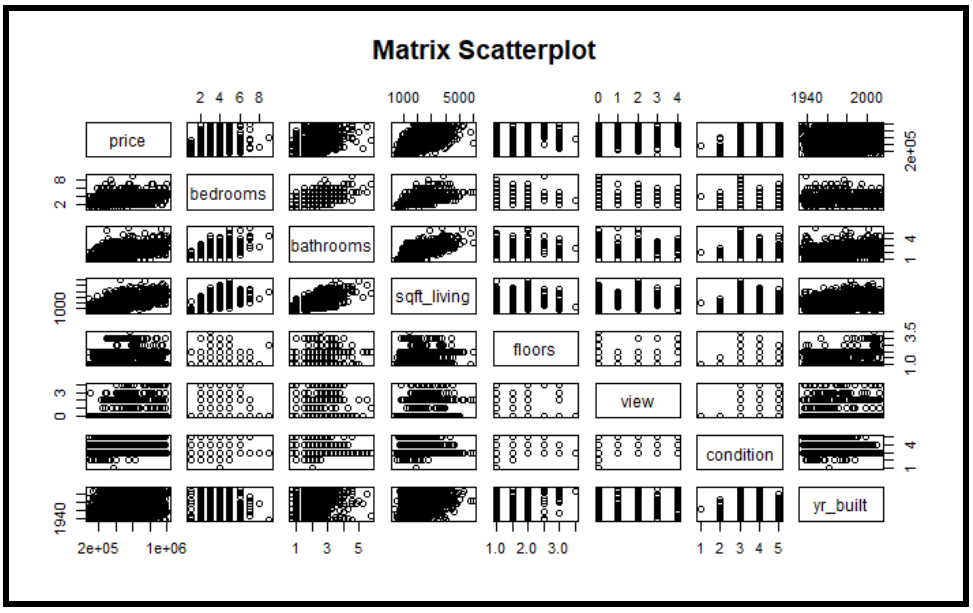

Fig.1: Matrix Scatterplot of price, bedrooms, bathrooms, sqft_living, floors, view, condition, sqft_above, and yr_built

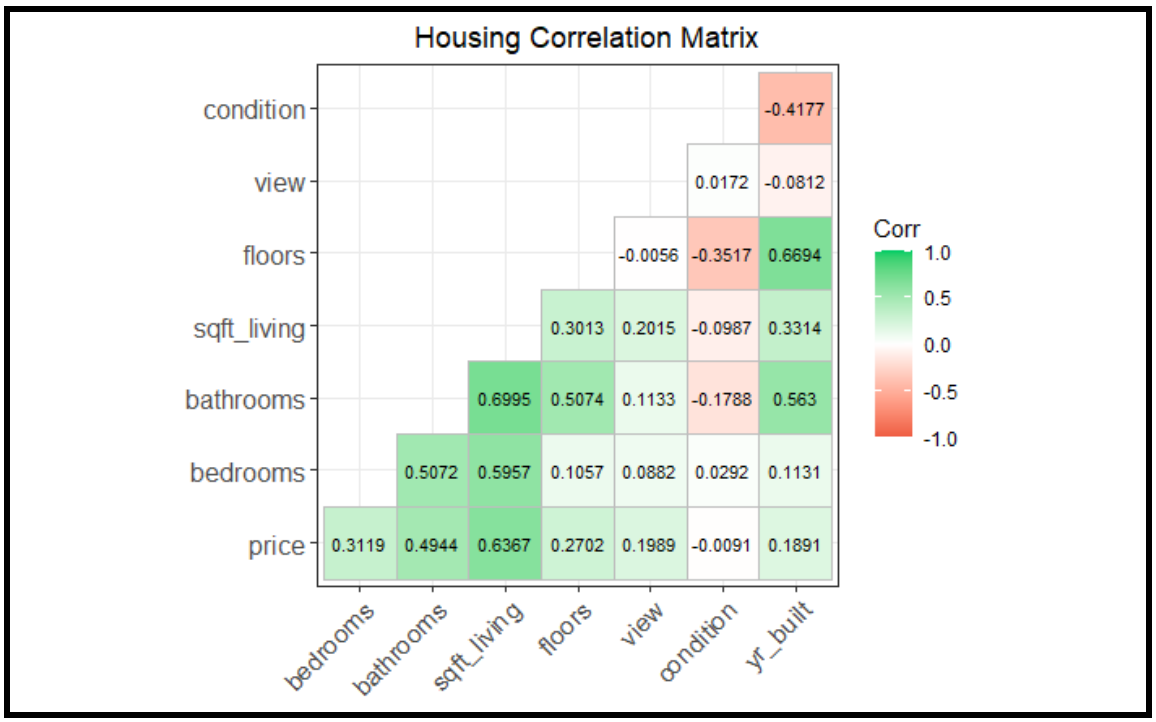

Fig.2: Color-Coded Correlation Matrix of price, bedrooms, bathrooms, sqft_living, floors, view, condition, sqft_above, and yr_built

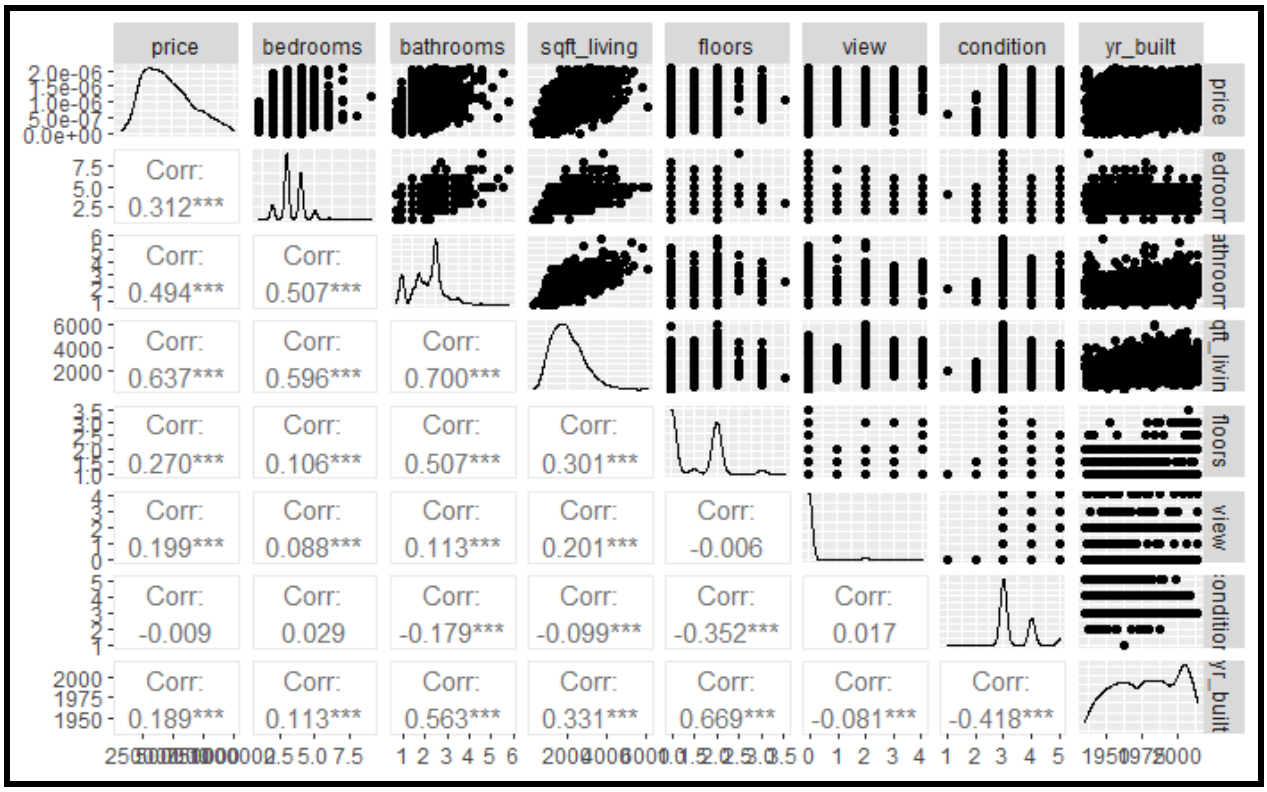

Fig.3: Pair Plot of Correlation Matrix and Points

The variable *sqft_living* looks the most strongly associated with *price* because the Pearson's correlation coefficient value, also known as the *r* value, between *sqft_living* and *price* is the highest at 0.597. After looking at the scatterplot of each variable versus *price*, it appears that all the variables are linear and we will not need to perform any transformations for the time being. Additionally, *sqft_living* and *bedrooms* have a high correlation of 0.6012, and *floors* and *yr_built* have a correlation of .4979. We may need to standardize or add an interaction variable to address a potential multicollinearity issue. Yes, the sign of the association between the predictors and price is as expected. However, this model may need to be adjusted due to multicollinearity issues which can be seen by an unusual coefficient sign. We also encountered unusual values that are far below and above reasonable house prices that may need to be filtered out of the data. There are unusual observations in our dataset. For example, there were two observations with high price values such as $26,590,000 and $12,899,000. Our dataset also has houses that are said to be sold for $0.

VI & VII. Variable Pre-Processing & Residual Analysis

	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront1	view	condition	yr_built	renovatedTRUE	basementTRUE	p	Cp	AIC	BIC	Rsq	Rsq.adj	s
1 (1)			*									2	207.982296	69341.83	69353.78	0.4054104	0.4052058	150580.6
1 (2)		*										2	1050.667900	70038.69	70050.64	0.2444031	0.2441430	169748.3
2 (1)	*		*									3	173.126117	69309.18	69327.11	0.4124523	0.4120478	149712.0
2 (2)			*		*							3	174.662679	69310.63	69328.56	0.4121587	0.4117540	149749.4
3 (1)			*		*				*			4	127.285585	69265.59	69289.49	0.4215930	0.4209954	148568.5
3 (2)			*		*			*				4	136.974706	69274.88	69298.78	0.4197417	0.4191423	148806.1
4 (1)	*		*		*				*			5	93.347501	69232.86	69262.73	0.4284595	0.4276719	147709.4
4 (2)	*	*	*						*			5	100.264705	69239.57	69269.45	0.4271376	0.4263485	147880.1
5 (1)	*	*	*		*				*			6	48.848145	69189.30	69225.15	0.4373438	0.4363744	146582.1
5 (2)	*		*		*			*	*			6	70.329954	69210.43	69246.29	0.4332394	0.4322629	147115.8
6 (1)	*	*	*		*			*	*			7	30.102350	69170.74	69212.57	0.4413076	0.4401521	146090.1
6 (2)	*	*	*		*		*		*			7	35.317440	69175.92	69217.75	0.4403112	0.4391536	146220.3
7 (1)	*	*	*		*		*	*	*			8	15.597897	69156.28	69204.08	0.4444610	0.4431201	145702.3
7 (2)	*	*	*		*			*	*		*	8	20.950879	69161.63	69209.43	0.4434383	0.4420948	145836.4
8 (1)	*	*	*		*		*	*	*		*	9	8.617513	69149.28	69203.06	0.4461769	0.4446486	145502.2
8 (2)	*	*	*		*		*	*	*	*		9	15.651144	69156.33	69210.11	0.4448330	0.4433010	145678.7

Fig.4: Best Subset Table

p	Cp	AIC	BIC	Rsq	Rsq.adj	s
9	8.617513	69149.28	69203.06	0.4461769	0.4446486	145502.2
8	15.597897	69156.28	69204.08	0.4444610	0.4431201	145702.3
9	15.651144	69156.33	69210.11	0.4448330	0.4433010	145678.7

Fig.5: Close-up of Best Subset Table

Another way we tried to determine the best fit for the model was through a variable selection technique known as *best subsets*. Using price as the response and the rest of the variables except for *date*, *street*, *city*, *statezip*, and *country*, we fitted a model called *fullModel* and ran the best subsets algorithm on it. The best subsets algorithm explores all possible models for each number of coefficients to determine the best models. From this, we looked for the single best model.

After running the best subsets algorithm, the best model we determined was that of predictors *bedrooms*, *bathrooms*, *sqft_living*, *floors*, *view*, *condition*, *yr_built*, and *basement* since the value of Mallow's C_p (8.6175) is approximately the same as p (9), the number of coefficients in the chosen model, which is desirable. 8(1) is the best model due to the smallest bias and variance.

Additionally, the selected model has the second-lowest value for Schwarz's Bayesian Information Criterion (BIC), as well as a comparable value for Akaike's Information Criterion (AIC). The adjusted R^2 value for the selected model is also the highest than those of the other model subsets'. This was actually the second time we ran this procedure. The first time, we were told to use sqft_basement, but there was a multicollinearity issue. To solve this problem, we decided to make a categorical variable, basement, on whether the house had a basement or not.

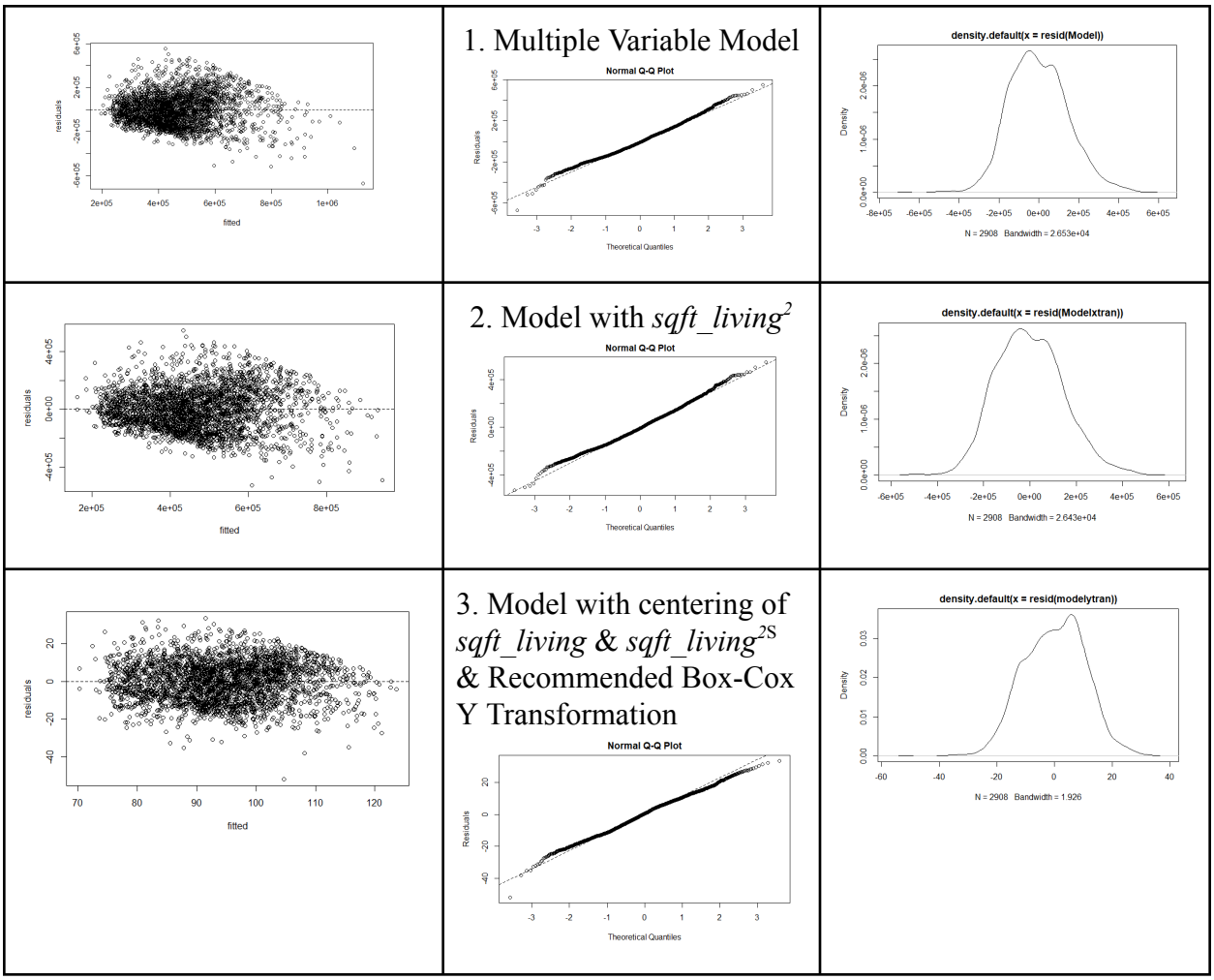

Fig.6: Transformation Steps

1. We decided to run a regression model with the best combination of variables from the best subset method in predicting housing prices. After doing so, we noticed that there was a structural issue in the model. Since the normality assumption was met and the equal variance assumption seemed borderline. We tried using the formal assumption tests, however, with the large sample sizes we had they would reject for the most minute deviations, making the tests unreliable. We looked at the individual residual vs predictor scatterplots and saw that sqft_living had a curved relationship with the residuals. Based on the results of the residual plots, we decided to add a polynomial term to x and test the model.

2. The new model seemed like an improvement from the previous, however, it also had issues that needed to be adjusted. The adjusted R^2 increased, and both the polynomial terms and other terms were significant, however, we noticed that the VIF indicated there was a multi-heteroscedasticity issue. In order to address that we decided to center $sqft\_living^2$ and sqft_living. The normality assumption seems met, however, the form and equal variance assumptions are violated. Looking at the residuals vs fitted plot, since equal variance seemed inappropriate, we conducted the box-cox procedure to determine what transformation to the response variable would be appropriate. From the box-cox plot as seen in figure <>, raising the response variable to the power of 0.35 looked like the best choice since it was within the 95% confidence interval of lambda.

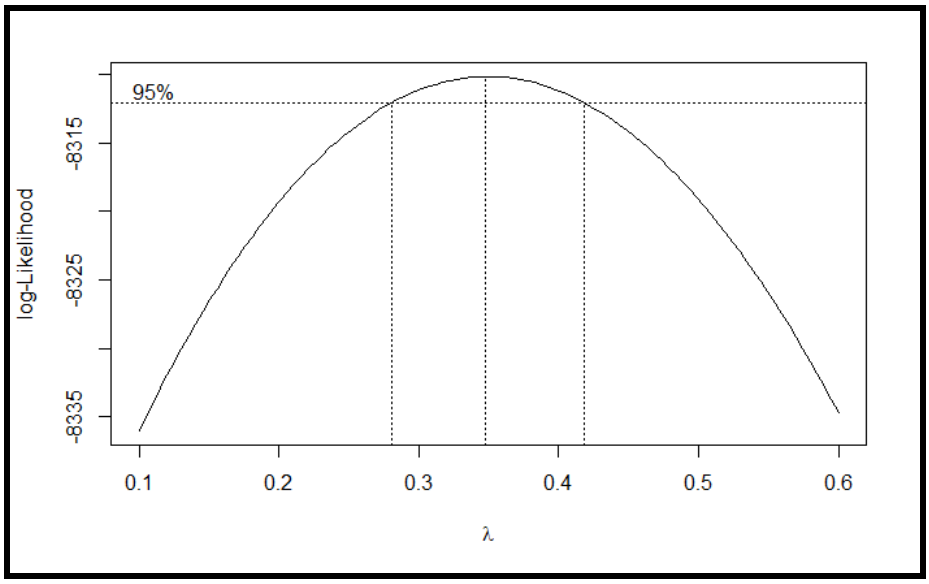

Fig. 7: Box-Cox Plot for Best Subset Model

3. The new model with the centering and price transformation did fix the equal variance and structure assumption, however, this was at the cost of the normality assumption. We learned that the Box-Cox method helps to address distributed data by transforming, however, there is no guarantee that data follow normality because it does not really check for normality. The Box-Cox method checks whether the standard deviation is the smallest or not. We decided that for our optimal model, we would need a transformation that decreases the power of price but is greater than the one recommended by the Box-Cox procedure.

We made this “balancing out adjustment” for our best model. Below are graphics that show that the methods and processes we used have created an accurate and appropriate regression model. Our filtration process was done in context and the lack of outlier and influential points indicate that the model is addressing the observations correctly.:

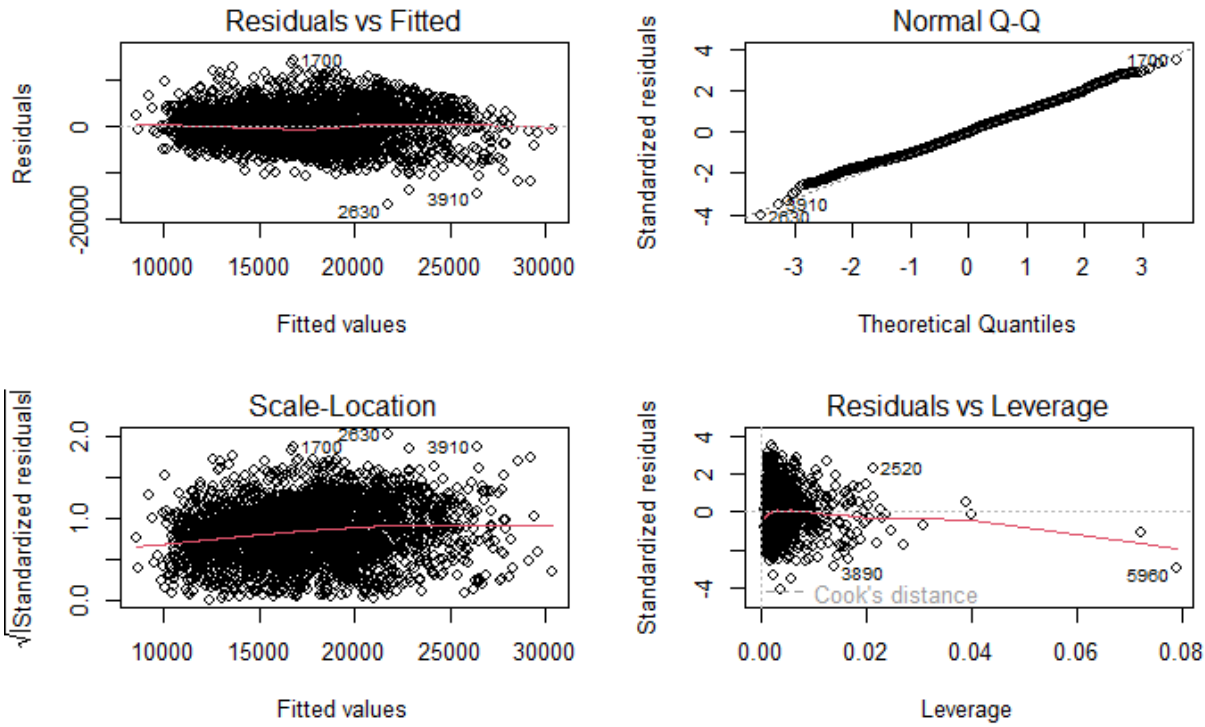

Fig.8: Residuals & Case Influence Diagnostics Plots for Best Model

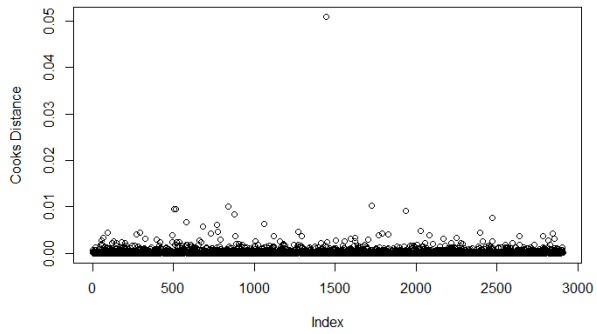

Fig.9: Leverage and Cook's Distance Plots for the Best Model

VIII. Fit a Linear Model

Coefficients	Estimate	Std. Error	t value	p values
Intercept	**8.887e+04**	**1.066e+04**	**8.335**	**< 2e-16**
bedrooms	**-1.018e+03**	**1.212e+02**	**-8.405**	**< 2e-16**
bathrooms	**1.054e+03**	**1.961e+02**	**5.377**	**8.19e-08**
I(sqft_living - mean(sqft_living))	**4.780e+00**	**1.742e-01**	**27.441**	**< 2e-16**
I((sqft_living - mean(sqft_living))^2)	**-5.433e-04**	**9.152e-05**	**-5.937**	**3.25e-09**
floors	**1.726e+03**	**2.028e+02**	**8.507**	**< 2e-16**
view	**5.273e+02**	**1.333e+02**	**3.956**	**7.81e-05**
condition	**6.058e+02**	**1.387e+02**	**4.368**	**1.30e-05**
yr_built	**-3.777e+01**	**5.402e+00**	**-6.992**	**3.34e-12**
basementTRUE	**5.276e+02**	**1.819e+02**	**2.901**	**0.00375**

Fig.10: Summary of Linear Model Fit of price, bedrooms, sqft_living, floors, waterfront, and yr_built

After looking at our best model, we identified two potential issues with the negative sign of the coefficients for *yr_built* and *bedrooms*. After doing additional research, we think that the coefficient of the number of bedrooms in the house (*bedrooms*) is negative due to the fact that an increase in the number of bedrooms in a house after adjusting for all other variables in the model, means that each bedroom is smaller(due to a constant size of the house in square feet (*sqft_living*)) and may contribute to a smaller predicted house price$^{0.75}$ and a smaller predicted house price overall. Additionally, *yr_built* has a negative coefficient sign, but we are unsure why that is occurring. Perhaps it is due to higher quality materials being used on older houses. The negative coefficient on centered sqft_living^2 with the positive coefficient on sqft_living shows that there is a diminishing marginal utility of each additional square foot of the size of the house. Other than these three coefficients, all the other coefficients are positive which makes sense in

the context of our model. Our model can only be generalized to houses in Washington state built after 1934 with prices between $8,000 and $1,000,000.

```
Residual standard error: 4121 on 2898 degrees of freedom
Multiple R-squared:  0.4528,    Adjusted R-squared:  0.4511
F-statistic: 266.5 on 9 and 2898 DF,  p-value: < 2.2e-16
```

Fig.11: Best Model Utility Test Output

Our model with the variables *bedrooms*, *bathrooms*, *centered sqft_living*, *centered sqft_living*^{2}, *floors*, *view*, *condition*, *yr_built*, and *basementTRUE* explains 45.28% of the variability in *Predicted Price*$^{.75}$. The average residual in our model is 4,121 units away from the predicted house *price*$^{.75}$.

Our interpretation for the intercept is with all of our predictors in our model to equal zero besides *sqft_living* which is equal to it's mean, the predicted house *price*$^{.75}$ is 88,870 units.

Our interpretation for the bathroom slope coefficient(quantitative) is after adjusting for all the other variables in the model, an increase in one additional bathroom increases the predicted house price$^{.75}$ by 1054 units.

Our interpretation for the basementTRUE slope coefficient(categorical) is after adjusting for all the other variables in the model, having a basement increases the predicted house price$^{.75}$ by 527.6 units.

Our interpretation for centered sqft_living^2 slope coefficient(polynomial) is after adjusting for all the other variables in the model, each one sqft_ living increases decreases the effect of each additional increase in sqft_living on price$^{.75}$ by -.0005433 . The negative coefficient on centered sqft_living^2 with the positive coefficient on sqft_living shows that there is a diminishing marginal utility of each additional square foot of the size of the house.Since all of our VIF values are below 5, we do not have any concerns with multicollinearity with our final model.

Predictors	VIF
bedrooms	1.758110
bathrooms	3.042700
I(sqft_living - mean(sqft_living))	3.122785
I((sqft_living - mean(sqft_living))^2)	1.412055
floors	2.184955
view	1.090994
condition	1.249379
yr_built	2.394639
basementTRUE	1.307500

Fig.12: VIF Table for Best Model

IX. Statistical Inference

In order to determine whether the combination of predictors we have chosen is significant or not, we perform an overall F-Test on the model. For the null hypothesis, we assume that none of the coefficients for the predictors of the prices of the houses are different from 0. For the alternative hypothesis, we state that at least one predictor coefficient is significantly different from 0, i.e, at least one predictor coefficient is not 0 and therefore significant in the model.

Overall F-Test for the model:

Model: $\text{price}^{0.75}$ ~ bedrooms + bathrooms + centered sqft_living + centered sqft_living^2 + floors + view + condition + yr_built + basementTRUE

H_0: $\beta_{\text{bedrooms}} = \beta_{\text{bathrooms}} = \beta_{\text{centered sqft\_living}} = \beta_{\text{centered sqft\_living\^2}} = \beta_{\text{floors}} = \beta_{\text{view}} = \beta_{\text{condition}} = \beta_{\text{yr\_built}} = \beta_{\text{basementTRUE}} = 0$

H_a: At least one $\beta_j \neq 0$

Test Statistic: F = 266.5 (on 9 and 2898 degrees of freedom)

P-value < 2.2e-16 < 0.001

Conclusion: Based on the small p-value (< 2.2e-16), we reject the null hypothesis and conclude there is extremely strong evidence that at least one of the parameters (β_j) is not 0.

Therefore, we find sufficient evidence to conclude that the model we selected contains significantly useful predictors of the price of houses in Washington.

Now, we want to ensure that there does not exist another much simpler model with a single parameter only that predicts the prices of houses more efficiently than our current model. To ensure this, we perform a partial F-Test on our selected model. In order to determine what sole predictor to compare our chosen model to, we looked at figure 2 to find the predictor with the highest correlation with predicted price, which happens to be *sqft_living*. Thus, we choose that to be the sole predictor of our reduced model.

Partial F-Test for the model, compared to centered sqft_living as the sole predictor:

Full Model: $\text{price}^{0.75}$ ~ bedrooms + bathrooms + centered sqft_living + centered sqft_living^2 + floors + view + condition + yr_built + basementTRUE

Reduced Model: $\text{price}^{0.75}$ ~ centered sqft_living

H_0: $\beta_{bedrooms} = \beta_{bathrooms} = \beta_{centered\ sqft\_living2} = \beta_{floors} = \beta_{view} = \beta_{condition} = \beta_{yr\_built} = \beta_{basementTRUE} = 0$

H_a: At least one of $\beta_j \neq 0$

Test Statistic: F = 32.831 (on 8 and 2898 degrees of freedom)

P-value < 2.2e-16 < 0.001

The null hypothesis states that all of the betas of the variables other than the square footage of the house have no statistically significant effect on the price$^{3/4}$ of a house after adjusting for the square footage of the house. The alternative hypothesis argues that at least one of the true effects on price$^{3/4}$ of a house after adjusting for the square footage of the house is statistically significant..

Conclusion: Based on the small p-value (< 2.2e-16), we reject the null hypothesis and conclude there is extremely strong evidence that at least one of the parameters (β_j) is not 0 i.e. our chosen model is more significant than the reduced model that uses the *centered sqft_living* as a sole predictor.

Based on the results of the partial F-Test, we find sufficient evidence that the number of bedrooms, bathrooms, floors, and the area of living squared, view rating, condition rating, year the house was built, and whether there is a basement or not are still significant predictors after adjusting for the area of the house (*centered sqft_living*).

The values in the intervals have been back-transformed from \$USD$^{0.75}$ to \$USD. As an interpretation of figure 13, we are at least 95% confident that the median house in the state of Washington has a predicted price$^{.75}$ between 18169 and 19991.87 units i.e. the predicted price is between \$477,648.08 and \$542589.30.

Additionally, we are 95% confident that a randomly selected house in Washington state with 3 bedrooms, 3 bathrooms, 2000 sqft_living, 1 floor, a 3 ranking for view, a 3 ranking for condition, a basement and built in 2000, has a predicted price$^{.75}$ between 10947.89 and 27212.99 units i.e. the predicted price is between \$243,093.81 and \$818,530.78.

We are at least 95% confident that the true beta coefficients of all predictors fall within each respective confidence interval.

Interval Type	Fit	Lower Bound(\$USD)	Upper Bound(\$USD)
Confidence	19080.44	\$477,648.08	\$542589.30
Prediction	19080.44	\$243,093.81	\$818,530.78

Fig. 13: 95% Confidence and Prediction Intervals for the price of an average home in Washington

Confidence Level = 1-(1-.95)/9		
	0.278%	99.722%
Intercept	6.29e+04	1.22e+05
bedrooms	-1.37e+03	-6.97e+02
bathrooms	6.41e+02	1.71e+03
centered sqft_living	4.36e+00	5.32e+00
centered sqft_living2	-8.29e-04	-3.24e-04
floors	9.99e+02	2.06e+03
view	1.65e+02	9.32e+02
condition	2.18e+02	9.89e+02
yr_built	-5.43e+01	-2.44e+01
basementTRUE	1.709531e+02	8.841890e+02

Fig. 14 : Bonferroni Adjusted Confidence Intervals for Slope Coefficients

X. Model validation

```
Call:
lm(formula = I(price^0.75) ~ bedrooms + bathrooms + I(sqft_living -
    mean(sqft_living)) + I((sqft_living - mean(sqft_living))^2) +
    floors + view + condition + yr_built + basement, data = testdata)

Residuals:
     Min       1Q   Median       3Q      Max
-12127.3  -3057.8    195.8   2670.0  14616.7

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          1.025e+05  2.080e+04   4.926 1.04e-06 ***
bedrooms                            -8.719e+02  2.250e+02  -3.875 0.000116 ***
bathrooms                            8.736e+02  3.713e+02   2.353 0.018893 *
I(sqft_living - mean(sqft_living))   4.713e+00  3.426e-01  13.756  < 2e-16 ***
I((sqft_living - mean(sqft_living))^2) -3.553e-04  1.717e-04  -2.069 0.038890 *
floors                               1.872e+03  4.043e+02   4.632 4.30e-06 ***
view                                 9.708e+02  2.433e+02   3.990 7.28e-05 ***
condition                            1.677e+02  2.672e+02   0.628 0.530493
yr_built                            -4.396e+01  1.053e+01  -4.174 3.36e-05 ***
basementTRUE                        -4.524e+02  3.601e+02  -1.256 0.209401
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4023 on 716 degrees of freedom
Multiple R-squared:  0.4863,    Adjusted R-squared:  0.4799
F-statistic: 75.32 on 9 and 716 DF,  p-value: < 2.2e-16

[1] 15964514
[1] 16986547
```

Our MSPE is 15,964,514 and MSE is 16,986,547. We see that the typical prediction error for the test selling prices for houses is smaller than the typical prediction error for the training selling prices for houses. Since our MSPE is close to our MSE, we can conclude that our model performed well on our test data set.

XI. Conclusion

Based on the model we fit on the selling prices of houses in the state of Washington, USA, the selling prices of houses tends to depend on nine factors: the number of bedrooms, the number of bathrooms, the square footage of the house, the square footage of the house squares, the number of floors, the view, the condition, the year the house was built, and whether or not the house has a basement. Based on the sign and values of the coefficients, as the square footage of the house increases, the selling price of the house also tends to increase, holding other variables constant. Additionally, an increase in the number of bathrooms in a house also tends to increase the selling price of it, holding other variables constant.

Overall, our model is significant and explains almost half of the variability in predicted $price^{.75}$, which is useful. Also, we had a small residual standard error which is a sign of a strong model. However, our model is not perfect. One weakness of our model is that we have 9 predictors which makes interpreting our model difficult. Additionally, we had to make an unusual y transformation which also made the interpretation more challenging. We think another variable to look at in the future would be municipality as the location of the house affect prices. Given the dataset we believe this model would be useful in predicting the selling price of other houses that fall in the scope of our sample.

Furthermore, our chosen model is constrained to a particular place, Washington, USA, and time, which is the year 2014. Due to this, our model is most likely only generalizable to the selling price of houses in Washington in 2014. Thus, we do not recommend utilizing this model to predict the selling price of homes in different states and/or for different periods of time.

XII. Appendix

All variables are from a dataset on Kaggle about houses in Washington. The author also looks at prices of houses in different countries in his other datasets. You can find the data here (hyperlink). We created 2 TRUE/FALSE variables from the kaggle dataset : Basement and renovation.

- date: a string variable that tracks the date of the house (observation) sold
- price: a quantitative variable that represents the selling price for the house ($ dollars)
- bedrooms: a quantitative variable that accounts for the number of bedrooms in the house
- bathrooms: a quantitative variable that accounts for the number of bathrooms in the house
- sqft_living: a quantitative variable that represents the size of actual house (in ft)
- sqft_lot: a quantitative variable that represents the size of whole property (in ft)
- floors: a quantitative variable that accounts for the number of floors in the house
- Waterfront: a categorical variable that states whether the house is on a waterfront or not
- view: a categorical variable that ranks the view from the house (1-5 with higher numbers meaning a better view)
- condition: a categorical variable that ranks the condition of the house (1-5 with higher numbers meaning a better condition)
- sqft_above: a numerical variable that represents the house size above ground level (in ft)
- sqft_basement: a numerical variable that represents the size of basement of house (in ft)
- yr_built: a numeric variable that represents the year the house was build (year)
- yr_renovated: a numeric variable that represents the year the house was renovated (year)
- Street: a string variable that has the address of the house sold (address)
- city: a string variable that has the city of the house sold (city in Washington)
- statezip: a string variable that has the state and zipcode of the house (WA zipcode)
- country: a string variable that has the country of the house sold (USA)

XII. Reference Page

Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition | Journal of Political Economy: Vol 82, No 1. (2022). Journal of Political Economy. https://www.journals.uchicago.edu/doi/10.1086/260169

Manhertz, Treh. "The U.S. Housing Market Gained More Value in 2020 than in Any Year Since 2005." *Zillow*, 26 Jan. 2021, https://www.zillow.com/research/zillow-total-housing-value-2020-28704/.

"National Housing Act (1934)." *Living New Deal*, 19 Oct. 2020, livingnewdeal.org/glossary/national-housing-act-1934/.

ysthehurricane. (2021, September 21). *House price prediction using R programming*. Kaggle.com; Kaggle. https://www.kaggle.com/code/ysthehurricane/house-price-prediction-using-r-programming/data