What Drives Home Value at Different Price Points:
A quantile regression approach to modeling hedonic attributes

Ava Allen, Svetlana Doronkina, Allison English and Otto JeckerByrne

Abstract

This paper applies quantile regression to examine the relationship between hedonic housing characteristics and sale prices across different price segments in Ames, Iowa. Traditional ordinary least squares (OLS) regression is limited in capturing variations in how housing attributes influence prices at different points of the distribution. By focusing on the 25th, 50th, and 75th quantiles, we explore how key predictors, such as overall quality, house size, and garage area, affect lower, median, and higher-priced homes. Our findings suggest that certain variables, such as OverallQual and X1stFlrSF, consistently influence price across all quantiles, while others, like KitchenAbvGr and YearsSinceRemodel, have varying effects at different price points. Notably, interaction terms between NeighborhoodCluster and HouseStyleCollapsed reveal that the impact of neighborhood quality on price varies significantly based on house style, particularly for homes in more affordable neighborhoods. The results underscore the importance of location and structural attributes, providing valuable insights for developers, builders, and real estate professionals to optimize investment decisions based on buyer expectations across different price levels.

Introduction

This paper addresses the quantile regression model, with a focus on size-related variables, and its capacity to measure the relationship between hedonic characteristics and house prices across different price levels in real estate analysis. Hedonic characteristics include physical attributes (e.g. size, age, number of bedrooms and bathrooms), location characteristics (e.g. proximity to schools and public transport), neighborhood attributes (e.g. safety and aesthetic appeal), and environmental factors (e.g. scenic view, climate, air quality). This model is found to be more effective than the traditional OLS approach in measuring the effect of hedonic characteristics across different house price levels. The reason is that the OLS method calculates the average house price for specific values of the explanatory variables. While it can assess the significance of an explanatory variable, it does not reveal how the significance of a variable varies across different quantiles. For this study, quantile regression is applied to housing data from Ames, Iowa, to develop a hedonic house price model and examine the impact of various hedonic attributes on house prices at different levels.

Literature Review/Past Studies using QR/Limitations

The quantile regression model was introduced in 1978 by Koenker and Basset as a more flexible approach to modeling house prices at different levels. Quantile

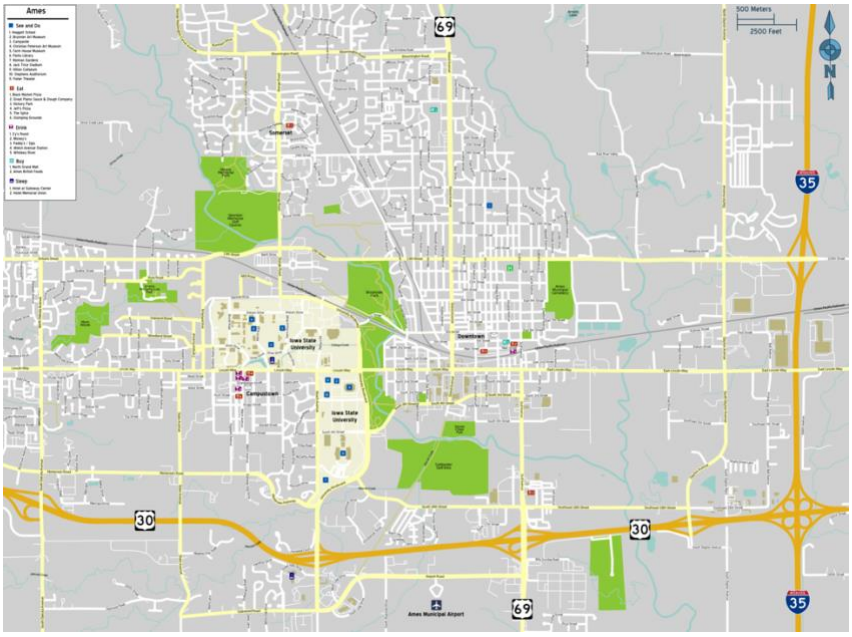
regression explains the determinants of the dependent variable at any point on the distribution of the dependent variable. Compared to OLS, this model allows for a more detailed picture of how various characteristics affect prices of houses. In 1999, Gyourko and Tracy found that the growth rate of quality in higher quality homes is much greater when modeled using quantile regression than previously estimated by Gyourko and Linneman in 1993 using the OLS method. Mak, Choy, and Ho (2010) used quantile regression to show that homebuyers' preferences for specific housing characteristics vary greatly across different quantiles in Hong Kong. In 2011, Ebru and Eban examined house prices in Istanbul using quantile regression and discovered that age, cable TV, heating system, garage, security, kitchen area, and number of rooms tend to increase house prices variably across different regions. In 2005, Lee, Chung, and Kim compared the differentiated effects of building age on house prices. This comparison showed apartment prices decreased until the buildings were 15-19 years and then began to rise again due to the prospect of redevelopment. Their article also highlights the limitations of using house price data projected from real estate agents instead of using real-transaction data.

The Kim et. al (2015) article notes on limitations associated with using data other than real-transaction data, such as court auction data. Kang and Liu (2014) implement quantile regression to investigate the impact that the 2008 financial crisis had on house prices in China and Taiwan. Interestingly, it was found that higher priced real estate was more affected by the financial crisis in Taiwan, while the opposite was true in China. The article by Kim et al. reaffirms the idea that buyers' preferences for certain features differ across price ranges. They found that proximity to metro stations and high schools is more significant for lower price quantiles, while scenic views have a larger positive impact on higher-priced homes. When discussing the varying results in the relationship between house prices and hedonic characteristics, the article lists potential reasons behind this variability. These reasons include the fact that each result is specific to its individual market of study, and that housing attributes vary in significance across different points on the conditional distribution of house prices.

Data

This study uses data from Ames, Iowa collected by the Ames Accessor's Office regarding residential properties that were sold between 2006-2010.

Figure 1. Map of Ames, Iowa



Ames is in Story County in central Iowa and is the home of Iowa State University (ISU). The presence of the college has a large effect on the city's economy, culture, and demographics. The city contains upper-class suburban neighborhoods, such as Northridge Heights and Stone Brook, as well as lower-class areas such as Meadow Village and Briardale. There is also an abundance of student housing, particularly in College Creek and South & West of Iowa State University. The smaller homes in Ames tend to be older and located in more affordable neighborhoods. Large luxury homes in Ames often exceed 3,000 square feet with 4-5 bedrooms, larger yards, and upscale finishes. This city has seen a growth in population over recent years and in turn an increase in housing market activity.

The Ames housing dataset is popular in machine learning and data analysis because of its real-world relevance, complexity, and large number of features. For applying quantile regression, the data is diverse with significant variation across price levels, making it ideal for capturing how different factors affect price level. The original dataset contains 1460 observations and 81 variables. The target variable is SalePrice. The data includes a number of general variables such as MSSubClass, the type of dwelling (e.g. 1-story, 2-story) and style (e.g. 1945 & older, duplex), LotArea which is the lot size in square feet, Neighborhood (where the house is located), Street which is the type of street access (e.g. paved, gravel), and 2 conditions which pertain to proximity to major roads and railroads. The data also includes variables on building and interior features. These include but are not limited to: YearBuilt, OverallQual (quality), Foundation (type of foundation), 1stFlrSF (first floor square footage), Heating, Electrical (Electrical system). GrLivArea (Above-grade ground living area in square feet), and several basement related features. There are also features describing bathrooms and bedrooms (number, type, quality), garage and parking (size, type, quality), and various amenities such as pools, fireplaces, fencing, decks, and porches. Sales features (other than price) include the month sold, year sold, and sale type.

Looking at the data as a whole, we can provide a brief summary of the relationship between sale price and various explanatory variables. Table 1 shows

descriptive statistics for the neighborhoods in Ames sorted by highest to lowest median house price. The statistics include average, median, minimum, maximum, and standard deviation for price, age, and size.

Figure 2. Descriptive Statistics

Neighborhood <chr>	AvgPrice <dbl>	MedianPrice <dbl>	MinPrice <int>	MaxPrice <int>	SDPrice <dbl>	AvgAge <dbl>	MedianAge <dbl>	MinAge <int>	MaxAge <int>	SDAge <dbl>	AvgSize <dbl>	MedianSize <dbl>	MinSize <int>	MaxSize <int>	SDSSize <dbl>
NridgHt	316270.62	315000	154000	611657	96392.545	2.142857	1.0	0	7	1.8898224	3831.896	3700	2612	6280	854.4561
NoRidge	335295.32	301500	190000	755000	121412.659	12.390244	12.0	6	18	2.7737885	5017.707	4836	3248	8952	1184.1272
StoneBr	310499.00	278000	170000	556581	112969.677	9.440000	7.0	0	25	9.0970691	3758.160	3484	2176	6558	1141.4580
Timber	242247.45	228475	137500	378500	64845.652	14.973684	8.0	0	62	17.0887775	3498.737	3379	2274	5796	783.1522
Somerst	225379.84	225500	144152	423000	56177.556	2.755814	1.0	0	11	3.2283161	3190.512	3129	2204	5220	593.7707
Veenker	238772.73	218000	162500	385000	72369.318	24.636364	29.0	11	33	8.1764629	3079.273	2874	2416	4334	627.9690
Crawfor	210624.73	200624	90350	392500	68866.395	65.941176	69.0	0	97	23.3087209	3575.686	3434	1388	6894	1079.6179
ClearCr	212565.43	200250	130000	328000	50231.539	41.214286	41.5	11	100	19.3761158	3559.107	3476	1976	5693	888.4586
CollgCr	197965.77	197200	110000	424870	51403.666	9.920000	6.0	0	37	10.6358391	2961.000	3000	1536	5656	861.8649
Blmngtn	194870.88	191000	159895	264561	30393.229	2.235294	2.0	0	6	1.6404806	2855.882	3000	2290	3138	276.3562
NWAmes	189050.07	182900	82500	299800	37172.218	32.191781	32.0	8	47	6.7672773	3455.562	3328	2064	5744	832.7204
Gilbert	192854.51	181000	141000	377500	35986.779	9.303797	8.0	0	59	10.7218378	3281.962	3186	1728	4924	593.8303
SawyerW	186555.80	179900	76000	320000	55651.998	19.661017	17.0	1	81	13.3695616	3190.034	3206	1504	6444	1015.9768
Mitchel	156270.12	153500	84500	271000	36486.625	26.000000	26.0	0	68	14.7761068	2588.408	2408	1536	4298	755.1620
NPkVill	142694.44	146000	127500	155000	9377.315	32.222222	32.0	31	35	1.3944334	2506.000	2644	1916	3096	475.0316
NAmes	145847.08	140000	87500	345000	33075.345	47.871111	48.0	9	89	8.8578210	2617.849	2392	1534	5956	820.0032
SWISU	142591.36	139500	60000	200000	32622.918	82.920000	83.0	66	97	9.3314879	3490.480	3382	876	6275	1181.0212
Blueste	137500.00	137500	124000	151000	19091.883	28.500000	28.5	28	29	0.7071068	2785.000	2785	2458	3112	462.4478
Sawyer	136793.14	135000	62383	190000	22345.129	44.216216	43.0	27	135	14.2110031	2430.270	2212	1344	5240	689.2498
BrkSide	124834.05	124300	39300	223500	40348.689	76.241379	78.0	36	99	12.0033772	2397.241	2421	668	4268	777.2521
Edwards	128219.70	121750	58500	320000	43208.616	51.970000	53.5	0	108	29.0751221	2680.080	2400	1210	11284	1310.4198
OldTown	128225.30	119000	37900	475000	52650.583	84.902655	86.0	5	136	23.0875525	2938.832	2736	1382	6986	1132.0008
BrDale	104493.75	106000	83000	125000	14330.176	36.250000	37.0	33	39	1.8797163	2286.375	2310	1974	2730	311.6573
IDOTRR	100123.78	103000	34900	169500	33376.710	79.810811	84.0	47	110	13.3932774	2254.676	2246	960	3636	565.2522
MeadowV	98576.47	88000	75000	151400	23491.050	35.470588	36.0	29	39	3.2233067	2117.882	2184	1260	5042	945.2517

25 rows | 1–16 of 21 columns

Based on the data, Northridge Heights has the highest median price while Northridge has the highest sale price. Northridge also has the greatest average house size, where "size" refers to the total interior living area. Additionally, the age of homes with higher median prices is relatively low when compared to other homes. Looking at lower priced homes, we see that size is lower while age increases. From these descriptive statistics, we can ascertain that house size and age are inversely related for higher priced homes. Neighborhoods such as College Creek and South & West of Iowa State University contain an abundance of college housing. These neighborhoods fall under the upper-middle and lower-middle range of average house prices and have higher average sizes, presumably to accommodate for multiple students.

Model Specification

$$\begin{aligned} \boldsymbol{SalePrice} = {} & \boldsymbol{\beta 0} + \boldsymbol{\beta 1} * \boldsymbol{LotFrontage} + \boldsymbol{\beta 2} * \boldsymbol{OverallQual} + \boldsymbol{\beta 3} * \boldsymbol{X1stFlrSf} + \boldsymbol{\beta 4} \\ & * \boldsymbol{KitchenAbvGr} + \boldsymbol{\beta 5} * \boldsymbol{Fireplaces} + \boldsymbol{\beta 6} * \boldsymbol{GarageArea} + \boldsymbol{\beta 7} \\ & * \boldsymbol{WoodDeckSF} + \boldsymbol{\beta 8} * \boldsymbol{OpenPorchSF} + \boldsymbol{\beta 9} * \boldsymbol{Age} + \boldsymbol{\beta 10} \\ & * \boldsymbol{YearsSinceRemodel} + \boldsymbol{\gamma 1} * \boldsymbol{FlrInteraction} + \boldsymbol{\delta 1} \\ & * \boldsymbol{NeighborhoodCluster} + \boldsymbol{\delta 2} * \boldsymbol{HouseStyleCollapsed} + \boldsymbol{\delta 3} \\ & * (\boldsymbol{NeighborhoodCluster} * \boldsymbol{HouseStyleCollapsed}) + \boldsymbol{\varepsilon} \end{aligned}$$

This equation represents the final hedonic pricing model used in this study. For this model, we focused on mostly size related variables to see how they affect sale price, as these variables provide strong indicators of value. This is supported by the fact that larger homes generally tend to be priced higher. The model also includes key property characteristics and specific significant attributes. LotFrontage represents the length of the street connected to the property in feet. OverallQual represents overall material and

finish quality. 1stFlrSf is the first-floor square footage. The KitchenAbvGr variable represents kitchens located above ground level. Fireplaces is the number of fireplaces, and although it is not size related, it was found to be consistently positive and significant in the model's performance. GarageArea contains the size of garage in square feet. WoodDeckSF and OpenPorchSF represent wood deck and open porch area in square feet. Age is a variable created by subtracting YearBuilt from YrSold. Similarly, YearSinceRemodel was created by subtracting YearRemodAdd from YrSold. FlrInteraction is an interaction term which equals X1stFlrS times X2ndFlrS. NeighborhoodCluster is a categorical variable created using k-means clustering which groups neighborhoods based on their grade or rating according to external "neighborhood_ratings" data, obtained from Niche.com. The data contains letter grades for each neighborhood that represent different characteristics of the neighborhoods such as housing, jobs, and schools. These letter grades are converted to numeric scores and used as inputs for k-means clustering. The algorithm then groups all the neighborhoods into one of three clusters: Cluster1 = high-rated, Cluster 2 = medium-rated, and Cluster3 = low-rated. HouseStyleCollapsed is an attribute which simplifies the classification of house styles by collapsing several styles into two broader groups, 1Story and 2Story. Finally, there is an interaction term between NeighborhoodCluster and HouseStyleCollapsed, allowing us to see the effect of neighborhood on sale price across different house styles.

- β = numeric main effect
- γ = numeric interaction
- δ = categorical main or interaction effect
- ε = error term, accounts for extraneous factors

After experimenting with multiple model specifications, we found this model to be the best balance between predictive accuracy and model complexity. The inclusion of both individual predictors and interaction terms allows for a comprehensive picture of the relationship between house characteristics and sale price. Categorical variables, NeighborhoodCluster and HouseStyleCollapsed, provide insight into how locational and structural attributes influence price as well.

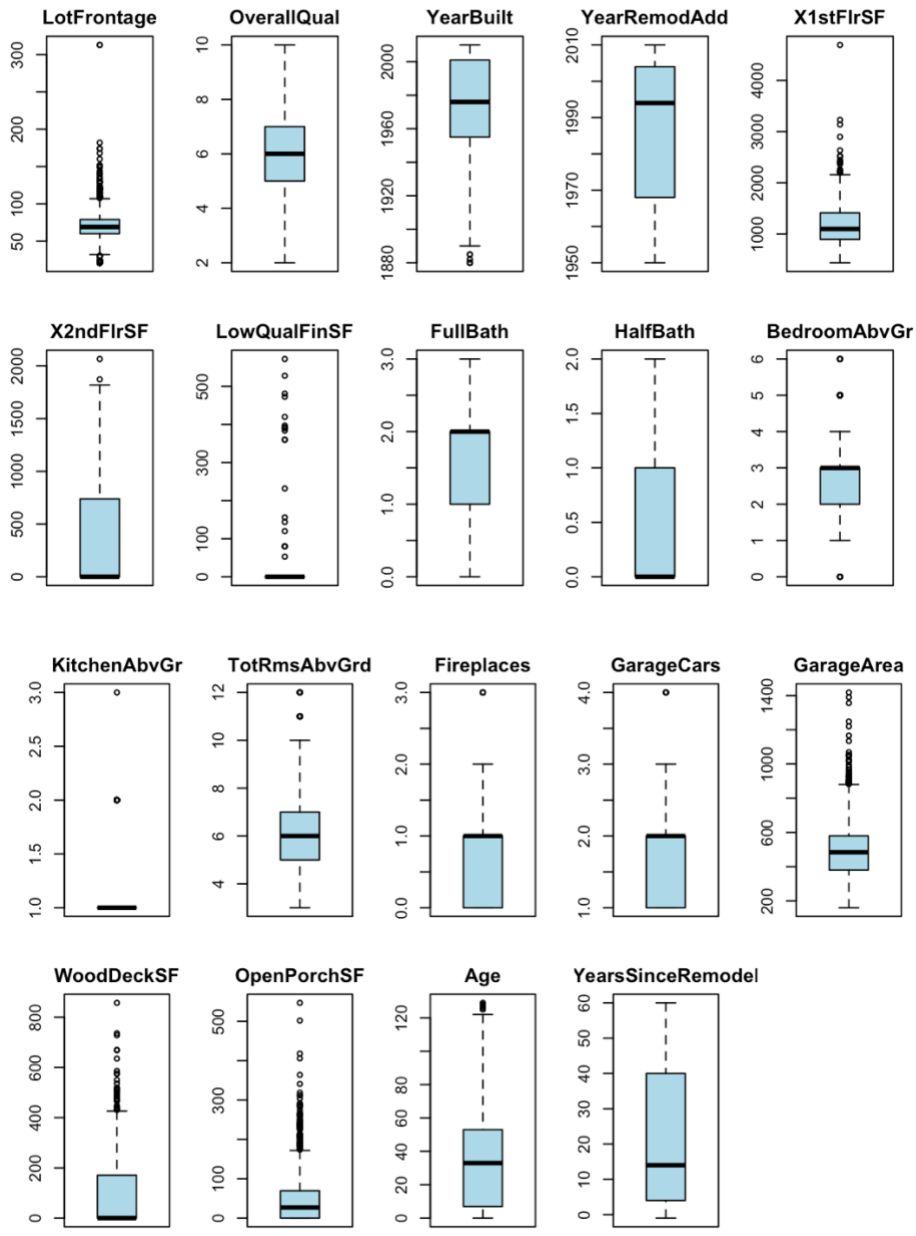
Methodology

As previously mentioned, this dataset originates from the Ames Housing training subset, which is competition data has 1460 instances across 81 different features. An external dataset was used to provide neighborhood ratings, which we transformed into scores that we used for clustering neighborhoods using k-means. A quantile regression model was developed to help estimate how housing features will impact sale prices across three different quantiles (25^{th}, 50^{th}, and 75^{th}), representing lower, middle, and higher end priced homes.

Initially, for data cleaning and variable selection, the top 30 most relevant features were selected based on the correlation to our target variable. Two new numeric variables were developed for simplicity, Age was calculated using YrSold – YearBuilt,

and YearSinceRemodel = YrSold - YearRemodAdd. These two new variables more accurately represent the age of the house and the year since the remodel. For handling missing data, the missing values for LotFrontage were replaced with the median, rather than the mean to lessen outlier influence. Any GarageType rows with missing values were eliminated entirely to maintain consistency for the categorical encoding since it is a categorical variable. For outlier detection, boxplots were used to visually identify outliers for numeric variables. Outliers were removed only if they were deemed implausible and likely to be errors, while others were kept if they appeared to be realistic for larger homes.

Figure 3. Boxplot Visualizations



In terms of categorical encoding, MSZoning, LotShape, Neighborhood, ExterQual, Kitchen Qual, and GarageType were converted into factors so that they could be easily integrated into the regression. HouseStyle, however, was simplified into two factors (1Story, 2Story) for simplicity and to reduce data dimensionality. Additionally, interaction terms were created to express key relationships between specific variables. These five interactions included Flr_Interaction = X1stFlrSF × X2ndFlrSF, Room_Interaction = TotRmsAbvGrd × BedroomAbvGr, Bath_Interaction = FullBath × HalfBath, Garage_Interaction = GarageCars × encoded GarageType, and Remodel_Interaction = YearBuilt × YearRemodAdd; which were developed to more effectively model the joint effects of variables. For neighborhood clustering, the external file neighborhood_ratings.xlsx was imported and then mapped from a letter grade system into numeric scores, then standardized. This integration allowed neighborhood quality to be included as a quantitative feature in the model. The new NeighborhoodCluster label was integrated back into the dataset and used as a categorical variable. Continuing to the model construction, multiple models were built sequentially until our final model was developed.

Model 1: Full model with main effects + numeric interaction terms

$$\begin{aligned}\text{Sale Price}_\tau = \ & \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{OverallQual} + \beta_3 \text{YearBuilt} + \beta_4 \text{YearRemodAdd} \\ & + \beta_5 \text{X1stFlrSF} + \beta_6 \text{X2ndFlrSF} + \beta_7 \text{LowQualFinSF} + \beta_8 \text{FullBath} + \beta_9 \text{HalfBath} \\ & + \beta_{10} \text{BedroomAbvGr} + \beta_{11} \text{KitchenAbvGr} + \beta_{12} \text{TotRmsAbvGrd} + \beta_{13} \text{Fireplaces} \\ & + \beta_{14} \text{GarageArea} + \beta_{15} \text{WoodDeckSF} + \beta_{16} \text{OpenPorchSF} \\ & + \gamma_1 \text{Flr\_Interaction} + \gamma_2 \text{Room\_Interaction} + \gamma_3 \text{Bath\_Interaction} \\ & + \gamma_4 \text{Garage\_Interaction} + \gamma_5 \text{Remodel\_Interaction} + \epsilon_\tau\end{aligned}$$

Model 2: Reduced model with key predictors + categorical main effects

$$\begin{aligned}\text{Sale Price}_\tau = \ & \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{OverallQual} + \beta_3 \text{YearBuilt} + \beta_4 \text{YearRemodAdd} \\ & + \beta_5 \text{X1stFlrSF} + \beta_6 \text{FullBath} + \beta_7 \text{KitchenAbvGr} + \beta_8 \text{Fireplaces} \\ & + \beta_9 \text{GarageArea} + \beta_{10} \text{WoodDeckSF} + \beta_{11} \text{OpenPorchSF} \\ & + \gamma_1 \text{Flr\_Interaction} + \gamma_2 \text{Remodel\_Interaction} \\ & + \delta_1 \text{MSSubClass} + \delta_2 \text{Neighborhood} + \delta_3 \text{MSZoning} + \delta_4 \text{HouseStyle} + \epsilon_\tau\end{aligned}$$

Model 3: Reduction + modification + neighborhood categorical variable

$$\begin{aligned}\text{Sale Price}_\tau = {} & \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{OverallQual} + \beta_3 \text{X1stFlrSF} \\ & + \beta_4 \text{KitchenAbvGr} + \beta_5 \text{Fireplaces} + \beta_6 \text{GarageArea} \\ & + \beta_7 \text{WoodDeckSF} + \beta_8 \text{OpenPorchSF} + \beta_9 \text{Age} + \beta_{10} \text{YearsSinceRemodel} \\ & + \gamma_1 \text{Flr\_Interaction} + \delta_1 \text{Neighborhood} + \epsilon_\tau\end{aligned}$$

Model 4: Adding categorical interaction terms

$$\begin{aligned}\text{Sale Price}_\tau = {} & \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{OverallQual} + \beta_3 \text{X1stFlrSF} \\ & + \beta_4 \text{KitchenAbvGr} + \beta_5 \text{Fireplaces} + \beta_6 \text{GarageArea} \\ & + \beta_7 \text{WoodDeckSF} + \beta_8 \text{OpenPorchSF} + \beta_9 \text{Age} + \beta_{10} \text{YearsSinceRemodel} \\ & + \gamma_1 \text{Flr\_Interaction} \\ & + \delta_1 (\text{NeighborhoodCluster} * \text{HouseStyleCollapsed}) + \epsilon_\tau\end{aligned}$$

Model 5: Adding categorical terms as main effects (No interaction)

$$\begin{aligned}\text{Sale Price}_\tau = {} & \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{OverallQual} + \beta_3 \text{X1stFlrSF} \\ & + \beta_4 \text{KitchenAbvGr} + \beta_5 \text{Fireplaces} + \beta_6 \text{GarageArea} \\ & + \beta_7 \text{WoodDeckSF} + \beta_8 \text{OpenPorchSF} + \beta_9 \text{Age} + \beta_{10} \text{YearsSinceRemodel} \\ & + \gamma_1 \text{Flr\_Interaction} \\ & + \delta_1 \text{NeighborhoodCluster} + \delta_2 \text{HouseStyleCollapsed} + \epsilon_\tau\end{aligned}$$

Model 6: Categorical Variables (No baseline Levels)

$$\begin{aligned}\text{Sale Price}_\tau = {} & \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{OverallQual} + \beta_3 \text{X1stFlrSf} + \beta_4 \text{KitchenAbvGr} \\ & + \beta_5 \text{Fireplaces} + \beta_6 \text{GarageArea} + \beta_7 \text{WoodDeck} + \beta_8 \text{OpenPorchSF} \\ & + \beta_9 \text{Age} + \beta_{10} \text{YearSinceRemodel} + \gamma_1 \text{FlrInteraction} \\ & + \delta_1 \text{NeighborhoodCluster} + \delta_2 \text{HouseStyleCollapsed} \\ & + \delta_3 (\text{NeighborhoodCluster} * \text{HouseStyleCollapsed}) + \epsilon_\tau\end{aligned}$$

To support the rationale behind our modeling choices and methodology, analysis revealed that quantile regression is ideal for understanding how predictors have the most influence at different price points (i.e. across the three different quantiles). It models the conditional distribution of the response variable, not just the mean. For this case, it determines what matters more for cheaper versus more expensive homes. Regarding location effects, neighborhood clustering was used instead of addressing the neighborhood variable as a fixed effect. Lastly, interaction terms were developed and employed in the models to capture non-linear and multiplicative effects (how two variables together can impact sale price). Overall, these methods create a model that more accurately captures the complex and multifaceted determinants of housing prices.

Empirical Results

Figures 4, 5, and 6 display the coefficients, standard errors, and statistical significance of each predictor variable and interaction in the model across the three quantiles (25th, 50th, and 75th). These results will be used to conduct an empirical analysis on the impact of the explanatory variables at varying points of the price-level distribution.

Figure 4. Results for Quantile Regression at τ = 0.25

```
Call: rq(formula = y_vector ~ X_matrix - 1, tau = tau_val)

tau: [1] 0.25

Coefficients:
```

	Value	Std. Error	t value	Pr(>|t|)
X_matrixLotFrontage	259.33574	54.14920	4.78928	0.00000
X_matrixOverallQual	12196.08800	1040.47774	11.72162	0.00000
X_matrixX1stFlrSF	49.61872	4.17993	11.87071	0.00000
X_matrixKitchenAbvGr	-30607.60279	4236.06893	-7.22547	0.00000
X_matrixFireplaces	4953.65004	1296.73820	3.82008	0.00014
X_matrixGarageArea	40.39888	6.54641	6.17115	0.00000
X_matrixWoodDeckSF	25.20290	7.71044	3.26867	0.00111
X_matrixOpenPorchSF	28.13457	13.52243	2.08059	0.03766
X_matrixAge	-440.81529	47.33668	-9.31234	0.00000
X_matrixYearsSinceRemodel	-188.98776	55.25102	-3.42053	0.00064
X_matrixFlr_Interaction	0.03376	0.00478	7.06218	0.00000
X_matrixNeighborhoodCluster1:HouseStyleCollapsed1Story	29678.52433	10395.34077	2.85498	0.00437
X_matrixNeighborhoodCluster1:HouseStyleCollapsed2Story	35727.30644	9761.96357	3.65985	0.00026
X_matrixNeighborhoodCluster2:HouseStyleCollapsed1Story	24554.13317	9672.49747	2.53855	0.01124
X_matrixNeighborhoodCluster2:HouseStyleCollapsed2Story	25583.68905	10104.87262	2.53182	0.01146
X_matrixNeighborhoodCluster3:HouseStyleCollapsed1Story	20752.03066	9789.55151	2.11981	0.03420
X_matrixNeighborhoodCluster3:HouseStyleCollapsed2Story	23438.41775	9120.60204	2.56983	0.01028

Figure 5. Results for Quantile Regression at τ = 0.50

```
Call: rq(formula = y_vector ~ X_matrix - 1, tau = tau_val)

tau: [1] 0.5

Coefficients:
```

	Value	Std. Error	t value	Pr(>|t|)
X_matrixLotFrontage	305.97079	43.17544	7.08669	0.00000
X_matrixOverallQual	12895.95671	1082.22242	11.91618	0.00000
X_matrixX1stFlrSF	61.50589	4.09486	15.02028	0.00000
X_matrixKitchenAbvGr	-39013.36226	4750.36684	-8.21271	0.00000
X_matrixFireplaces	3928.28071	1674.24653	2.34630	0.01910
X_matrixGarageArea	35.40781	6.01751	5.88413	0.00000
X_matrixWoodDeckSF	24.02758	6.86274	3.50116	0.00048
X_matrixOpenPorchSF	26.51429	12.32517	2.15123	0.03163
X_matrixAge	-333.81478	44.18855	-7.55433	0.00000
X_matrixYearsSinceRemodel	-271.51103	48.69063	-5.57625	0.00000
X_matrixFlr_Interaction	0.04270	0.00449	9.49907	0.00000
X_matrixNeighborhoodCluster1:HouseStyleCollapsed1Story	32428.53571	8806.76373	3.68223	0.00024
X_matrixNeighborhoodCluster1:HouseStyleCollapsed2Story	35366.16811	9331.48682	3.78998	0.00016
X_matrixNeighborhoodCluster2:HouseStyleCollapsed1Story	24164.11526	8899.40951	2.71525	0.00671
X_matrixNeighborhoodCluster2:HouseStyleCollapsed2Story	22532.18897	8767.79379	2.56988	0.01028
X_matrixNeighborhoodCluster3:HouseStyleCollapsed1Story	25315.00465	9496.27290	2.66578	0.00777
X_matrixNeighborhoodCluster3:HouseStyleCollapsed2Story	21388.27888	8893.17414	2.40502	0.01630

Figure 6. Results for Quantile Regression at $\tau = 0.75$

```
Call: rq(formula = y_vector ~ X_matrix - 1, tau = tau_val)

tau: [1] 0.75

Coefficients:
```

	Value	Std. Error	t value	Pr(>|t|)
X_matrixLotFrontage	225.57429	71.62991	3.14916	0.00167
X_matrixOverallQual	14900.92933	1537.28048	9.69305	0.00000
X_matrixX1stFlrSF	77.58103	6.15339	12.60785	0.00000
X_matrixKitchenAbvGr	-34360.15157	8384.84499	-4.09789	0.00004
X_matrixFireplaces	5844.54981	2177.35653	2.68424	0.00736
X_matrixGarageArea	37.03869	7.35885	5.03321	0.00000
X_matrixWoodDeckSF	29.53329	7.71068	3.83018	0.00013
X_matrixOpenPorchSF	15.09952	18.00360	0.83869	0.40179
X_matrixAge	-354.24198	49.66062	-7.13326	0.00000
X_matrixYearsSinceRemodel	-263.81222	49.40378	-5.33992	0.00000
X_matrixFlr_Interaction	0.05805	0.00806	7.20103	0.00000
X_matrixNeighborhoodCluster1:HouseStyleCollapsed1Story	19060.64291	13522.89254	1.40951	0.15891
X_matrixNeighborhoodCluster1:HouseStyleCollapsed2Story	8119.58412	13092.44149	0.62017	0.53525
X_matrixNeighborhoodCluster2:HouseStyleCollapsed1Story	6941.19297	12128.37471	0.57231	0.56721
X_matrixNeighborhoodCluster2:HouseStyleCollapsed2Story	-3955.40594	12733.57045	-0.31063	0.75613
X_matrixNeighborhoodCluster3:HouseStyleCollapsed1Story	6356.14440	12319.94775	0.51592	0.60599
X_matrixNeighborhoodCluster3:HouseStyleCollapsed2Story	-4196.43102	12428.81158	-0.33764	0.73569

OverallQual is highly statistically significant and has a positive effect on sales price across all three quantiles. This demonstrates that higher overall quality equals higher sale price. X1stFlrInteraction also maintains a positive effect on sales price with values of 49.6187 (25th quantile), 61.5059 (50th quantile), and 77.5810 (75th quantile). The impact of this variable increases as the price of a home increases. Fireplaces has a higher positive effect in the 25th and 75th quantiles, with a greater effect in the 75th quantile than the 25th. Having multiple fireplaces is a luxury and thus will have a greater effect on the higher priced houses Age, KitchenAbvGr, and YearsSinceRemodel consistently have a negative effect on sale price. Older homes are less desirable due to the need for repairs, lack of amenities, and unfavorable layouts. The nature and timing of remodeling work can have a negative effect on sale prices if buyers don't see an increase in value from the work or their preferences have changed. Negative effects of KitchenAbvGr could be due to older homes with additional kitchens that are not seen as necessary to buyers. High correlation between variables could also be the cause. A correlation matrix of numeric predictors showed weak multicollinearity between KitchenAbvGr and other numeric variables. Interactions between neighborhood and house style have a greater positive effect on sale price in the 50th and 25th quantiles. In the 75th quantile, these interactions are statistically insignificant and potentially unfavorable for sales price. This suggests that other factors drive the price of high-end homes, such as those related to luxury and uniqueness. 2-story homes have a lesser and sometimes negative effect on price when compared to the 25th and 50th quantile because it is a buyer expectation for higher-priced homes and therefore is does not greatly impact value.

Conclusion/Business Implications/Limitations

This paper applies quantile regression as a method to model housing prices using a set of selected predictor variables, with a focus on size-related variables. After considerable data cleaning and the creation of interaction terms and clustering neighborhoods based on neighborhood quality data - the analysis explored how several

features influence different points in the housing price distribution. Six different quantile regression models were created, each one with different structures and complexity. We also included interaction effects for both numerical and categorical variables to represent more fully developed relationships among variables. The final model reveals that predictor variables like overall quality of the house, floor area, and garage size are all price determinants consistently across all price quantiles.

From a business perspective, our quantile regression model suggests that different predictors are more relevant at different price levels. For instance, higher-end buyers tend to value features such as deck size and updated models when compared to lower-end buyers. Additionally, the interaction between the NeighborhoodCluster and the HouseStyle variable (indicating 1-story, 2-story, split level, etc.) reveals a significant relationship between these two variables. Not all house styles have the same impact in each neighborhood. In higher-priced neighborhoods, the 2-story attribute does not dramatically increase the value, suggesting that it is a buyer expectation and therefore does not justify a higher price. However, in more affordable neighborhoods, a 2-story attribute appears to have a stronger impact on the value. This interaction highlights the importance of paying close attention to which features contribute the most value to homes based on which neighborhoods they're located in. Builders, house flippers, and designers can leverage this information to help improve marketing strategies or product offerings when working with consumer segments across the different quantiles. According to these results, real estate developers and homebuilders should prioritize modern upgrades – such as larger decks – when selling to higher-end buyers. Meanwhile, sellers looking to work in more conservatively priced neighborhoods should focus on house structure and design. This approach takes buyer expectations into consideration while optimizing spending.

Some possible limitations of this model include multicollinearity, which are common when employing a quantile regression model with predictors that are highly correlated with many interaction terms. In this case, multicollinearity can make it more difficult to compare results across different quantiles; and adding interaction terms only amplifies the issue. For real estate agents and developers, this can make it more difficult to determine the individual impact of each specific variable, making it harder to use the results to prioritize certain design choices. As a result, developers may struggle with decision-making when it comes to which features are appropriate to prioritize based on the price segment they are working with.

References

Ebru, C., & Eban, A. (2011). Determinants of house prices in Istanbul: A quantile regression approach. *QualQuant*, 45(2), 305–317.

Kang, H. H., & Liu, S.-B. (2014). The impact of the 2008 financial crisis on housing prices in China and Taiwan: A quantile regression analysis. *Economic Modelling*, 42, 356–362.

Kim, H., Park, S. W., Lee, S., & Xue, X. (2015). Determinants of house prices in Seoul: A quantile regression approach. *Pacific Rim Property Research Journal*, *21*(2), 91–113.

Koenker, R., & Bassett Jr., G. (1978, January). Regression Quantiles, *Econometrica*, 46(1), 33–50.
Lee, B. S., Chung, E.-C., & Kim, Y. H. (2005). Dwelling age, redevelopment, and housing prices: The case of apartment complexes in Seoul. *Journal of Real Estate Finance and Eco-nomics*, 30(1), 55–80.