House Prices in King County: A Regression Analysis

Abstract

Our paper seeks to investigate the factors that affect King County's house prices from a data set of houses sold in King County between 2014 and 2015. After processing the data into our desired structure, we first perform stepwise linear regression on logarithm of house price with quadric and intersection terms of some variables inserted. With this model, we found that variables including area, number of different rooms and floors, years the house exists make a difference to the house price. Adopting the similar regression process on logarithm of house price per square feet, we have an equally satisfying model with variable floors excluded. We further polish our analysis with regression by region and quantile regression, getting more detailed information about how some variables' effects change among different regions or price levels.

Part I: Introduction

Scholars have long been interested in the dynamics of house prices. The literature has emphasized macro factors including city size,[1] real income, long-run interest rates,[2] etc. However, these do not specify which factors will affect the house prices within a given county or region. Therefore, our research intends to study the household (micro) level of house price differences and seek to identify some key factors that affect the house prices.

We have obtained the data from Kaggle,[3] and the dataset contains information on house sales in King County, WA, from 2014-2015. It contains information on house prices, areas of living room, basement, etc. in square feet, and other related information. A detailed explanation of

[1] Capozza, Dennis R., et al. "Determinants of real house price dynamics." (2002). *NBER*. https://www.nber.org/papers/w9262

[2] Algieri, Bernardina. "House price determinants: Fundamentals and underlying factors." *Comparative Economic Studies* 55.2 (2013): 315-341. https://link.springer.com/article/10.1057/ces.2013.3

[3] Untitled. "House Sales in King County, USA". *Kaggle*. https://www.kaggle.com/harlfoxem/housesalesprediction/version/1

the variables can be found in Appendix I.[4] A few new variables have been introduced and are also specified in Appendix I. The analysis of this dataset has been performed in R.

Throughout the paper, we perform the exploratory data analysis (EDA) to gauge insights from the data. The results of the EDA are shown in Part II. Following the EDA, we conduct a series of regression analysis, during which we fit the interaction and quadratic terms to improve the model. The model fitting and its explanation are included in Part III. In Part IV, we explain how the quantile regression is used to study how factors affect the house prices of different ranges. In the same section, we perform the regression by region to study regional differences for the regression estimates in different zip codes. In Part V, we summarize the findings and discuss some of their implications.

Part II: Exploratory Data Analysis

The purpose of the EDA is to clean and prepare the data for later regression analysis, as well as to find possible correlations between predictors and the response variable. The EDA has been conducted in a way to understand the structure of the dataset, create and drop variables as needed, and generate visualizations and tables to identify correlations.

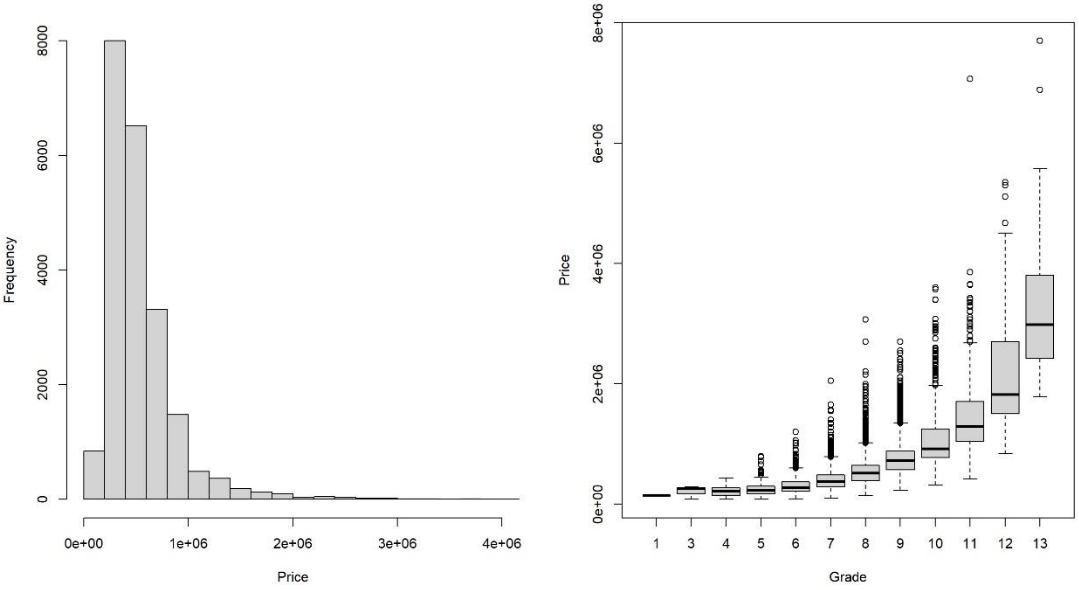
The dataset contains 21,613 observations and 23 variables. Most of the variables take numeric values and a few are string or Boolean variables. After understanding the data, a few other variables have been created: renovation (Boolean), yrs_present (numeric), and total_area_except_lot (numeric). Additionally, variables such as view and condition have been changed to factor variables, and zip code has been changed to character to represent different regions. A few variables have been dropped, including id and date, which will not be used, and sqft_living and sqft_lot, which are outdated and we will use sqft_living15 and sqft_lot15 instead.

After adjusting for the types of the variables, we investigate the distribution of the variables. The most important one is the response variable, price. It has a right-skewed bell-shaped distribution, where the center is around half a million but the maximum exceeds 7 million (Plot 1a). A log transformation is possibly ideal in the analysis. For the area variables, only the living area and the above area variables are approximately normally distributed. As for the basement and

[4] Some of the explanations are extracted from Rawat's analysis. Rawat, Neeraj Singh. "Data Analysis with Python: House Sales in King County, USA". https://courserasolution.blogspot.com/p/house-sales-in-king-county-usa.html

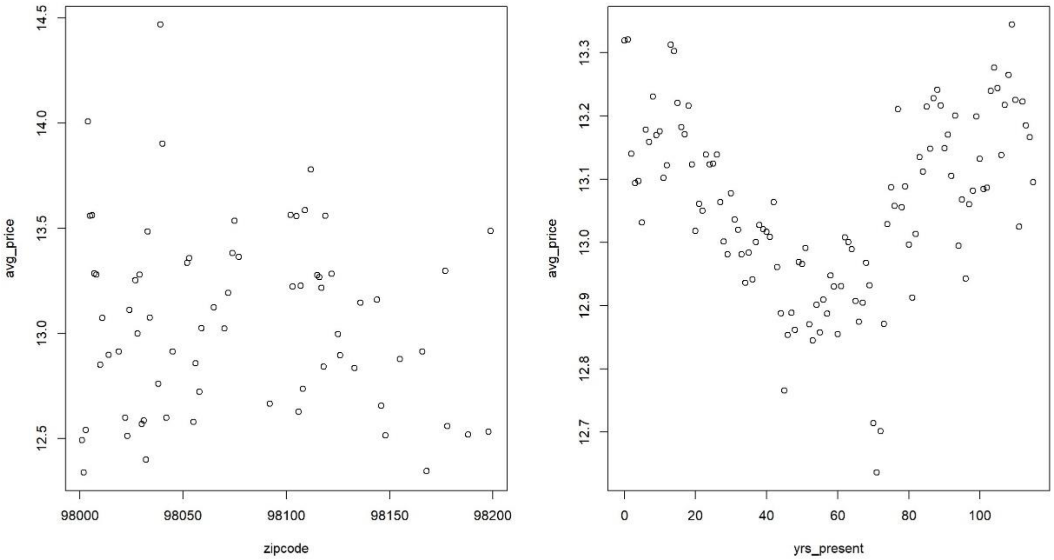
lot areas, the former has a lot of zero values and the latter has quite a few extreme values. Therefore, the basement and lot area variables will either not be included in the regression or used as levels per our needs. Sqft_lot15, for example, will be transformed to 5 levels as a categorical variable in the regression analysis.

The variables regarding the number of rooms and floors seem to behave well. And although view appears not to matter in the interaction plots, grade does make a difference (Plot 1b). Besides, there seems a clear difference in the mean price of houses in different zip codes (Plot 1c). Therefore, we have reason to hypothesize that the coefficient estimates in different regions may differ in magnitude, and it’s also possible that Simpson’s Paradox may occur. Moreover, there is a quadratic relationship between the price and the yrs_present variable (Plot 1d). Therefore, the quadratic term may be included in the regression.



a. Histogram of price

b. Boxplot of price by grade

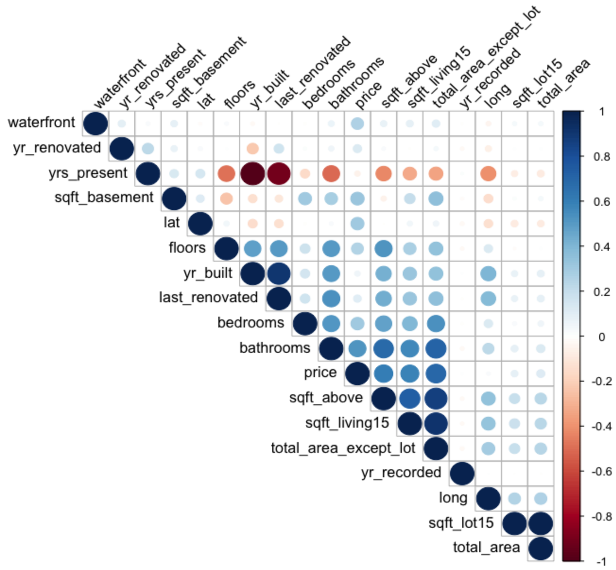


c. Average price by zipcode

d. Average price by years to present

Plot 1: Exploratory plots of data

Based on Pearson's correlation table in Plot 2, we can see that a few variables should be included in the regression, namely bedrooms, bathrooms, floors, sqft_above, sqft_basement, lat, sqft_living15, total_area_except_lot, waterfront. All these variables are correlated with the response variable, price. Also from the same table, we should be careful not to use some variables simultaneously due to concerns of collinearity. Two sets of variables exist: the first includes yr_built vs. last_renovated and yrs_present; the second includes total_area_except_lot vs. sqft_above and sqft_living15.



Plot 2: Correlation table of variables

EDA part has prepared the data for subsequent analysis and provides a roadmap for quantile and regional regressions.

Part III: Linear Regression

In this part, we will mainly use stepwise linear regression to explore the relationship between housing price and all explanatory factors.

Based on Pearson's correlation table, we include variables that have high correlation with housing price: sqft_above, lat, bedrooms, bathrooms, floors, sqft_basement, waterfront. Notice that among all data collected, 60.7% observations do not have any basement. This means if we include sqft_basement directly into the regression model, there might be some inaccuracy and difficulty in interpreting its coefficient. Thus, we transform sqft_basement into a Boolean variable of 0 and 1 (0: no basement; 1: has basement). Floors also appears a little wield. It is categorical, but has values such as 1.5 and 2.5, which is a bit difficult to interpret. Thus, we rescale it into a categorical between 1 to 6. As for sqft_living, though it is also correlated with housing price, we have to exclude it because it is highly correlated with sqft_above. And we can see that sqft_above indicates the total area of the house except the basement, which works just fine in the company of sqft_basement.

Besides these factors, we also add some extra explanatory variables for better fitting. First of all, common sense tells us that if a house has a better view or overall grading score, then it is definitely more expensive. Thus, we take grade, view into account. Second, houses of different ages have different facility conditions, which will definitely affect overall housing price. Renovation is a way to compensate for this problem, but it may also cause doubts of housing conditions from potential buyers. We include yrs_present alongside with renovation to reflect this factor. Lastly, although sqft_lot15 is almost uncorrelated with housing price, it is still part of total housing area. Thus, we include sqft_lot15 alongside sqft_above and sqft_basement. However, sqft_lot15 has a huge data range and many extreme values. If directly included into the regression model, this factor will possibly cause shrinkage of other regression coefficients, making them difficult to interpret. Thus, sqft_lot15 is transformed into a categorical variable with 5 levels based on its quantile. All factors included in the regression model, along with its type and any possible manipulation, are summarized in Appendix II.

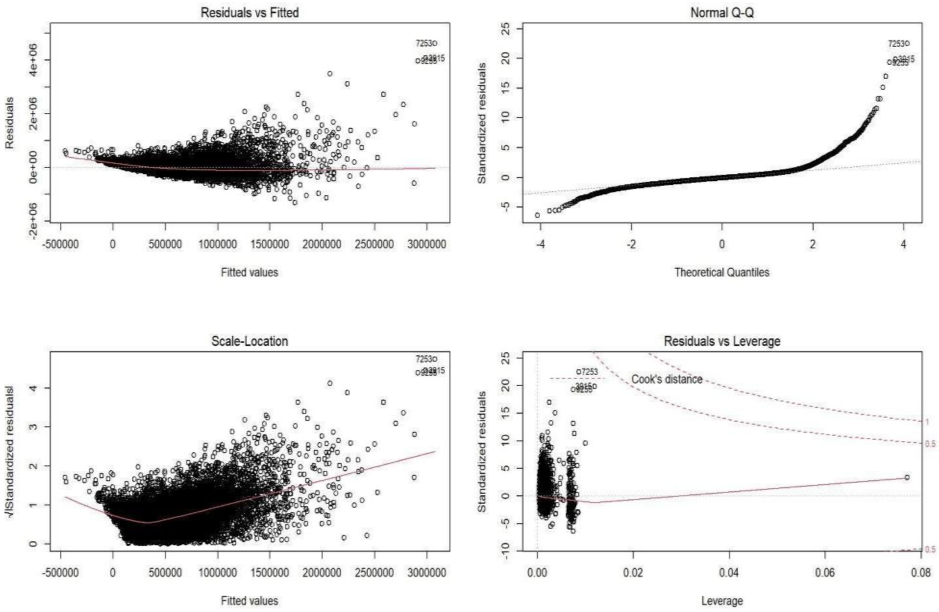
As we have discussed above, yrs_present represents house ages, which might be compensated by renovation. This indicates a possible interaction term: yrs_present*renovation. Furthermore, Plot 1d shows that the relationship between yrs_present and housing price is quadric. Thus, we should also include the quadratic term: yrs_present$^{\wedge 2}$ in the regression model.

To summarize, there are 12 variables in the model, as long as one interaction term and one quadratic term. For a regression problem, the potential model should be precise. However, it should also be as simple as possible to avoid overfitting problems and for explanation convenience. Thus, we adopt stepwise regression with a bidirectional elimination method for variable selection. This can also serve as a way to reduce collinearity between explanatory variables.

First of all, run stepwise regression on the following raw model.

$$\begin{aligned} price \sim sqft\_above + lat + yrs\_present + yrs\_present^{\wedge 2} + bedrooms + bathrooms \\ + floors + grade + view + sqft\_basement + sqft\_lot15 + renovation \\ + yrs\_present \times renovation + waterfront \end{aligned}$$

This raw model isn’t what we want. By plotting the below diagnosis plots, it is crystal clear that housing price has heteroskedasticity problem. Also, the Q-Q plot reveals the non-normality of regression residuals. These violations of classical linear regression assumptions indicate that the regression coefficients might not be as robust as what we desired.



Plot 3: Diagnosis plots of raw linear model

In order to deal with these violations, we introduce Box-Cox transformation. By taking log-transformation of housing price, variance of housing price will be stabilized. After this transformation, regression coefficients can be interpreted with regard to "average percentage change of housing price". New model and fitted coefficients are shown as below. Diagnosis plots are included in Appendix III.

$$
\begin{aligned}
log\ (price) \sim & sqft\_above + lat + yrs\_present + yrs\_present^{\wedge 2} + bedrooms + bathrooms \\
& + floors + grade + view + sqft\_basement + sqft\_lot15 + renovation \\
& + yrs\_present \times renovation + waterfront
\end{aligned}
$$

Variable	Estimate		Variable	Estimate	
(Intercept)	$1.121*10^{1}$	***[5]	view	$7.536*10^{-2}$	***
sqft_above	$1.996*10^{-4}$	***	sqft_basement1	$1.404*10^{-1}$	***
lat	$1.809*10^{-3}$	***	sqft_lot151	$-2.841*10^{-2}$	***
yrs_present	$1.115*10^{-3}$	***	sqft_lot152	$-1.256*10^{-1}$	***
yrs_present$^{\wedge 2}$	$2.351*10^{-5}$	***	sqft_lot153	$-8.766*10^{-2}$	***
bedrooms	$7.403*10^{-3}$	**	sqft_lot154	$-4.279*10^{-2}$	***
bathrooms	$7.169*10^{-2}$	***	waterfront1	$3.687*10^{-1}$	***

[5] Significance level: 0.001 (***); 0.01 (**); 0.05 (*)

floors	-1.437*10^{-2}	***	renovation1	3.120*10^{-1}	***
grade	1.798*10^{-1}	***	yrs_present*renovation1	-3.668*10^{-3}	***

Table 1: Regression coefficients on log(price)

After log-transformation of housing price, regression residuals look more like a normal distribution. All regression coefficients are significant. The adjusted R^2 is 0.7641, meaning that this model can explain 76.41% variation of housing price, which looks pretty good.

The coefficients of sqft_above, bedrooms, bathrooms, grade, view, sqft_basement_1, renovation_1, waterfront_1 are all positive. These results are pretty intuitive. For instance, given all other conditions, a house having more bathrooms is definitely more expensive, since more bathrooms requires more construction cost and bigger inside housing area. The interesting part is latitude. Coefficient of latitude is positive, which means given all other conditions, houses located north are more expensive than houses in the south. This sounds pretty weird. A possible explanation is that many affluent regions, like Shoreline, Lake Forest Park, and Clyde Hill, are all located in the northern part of King County. Rural regions near the Mount Rainier area, which belong to the southern part of King County, have relatively low housing prices compared to urban districts in the north.

Among all positive factors of housing price, grade, sqft_basement, renovation, waterfront have the biggest influence. This indicates three findings: 1. When dealing with house properties, people rely heavily on the house grading system. 2. People have a strong preference for houses with a view to the waterfront. 3. Houses with a basement or have been renovated worth more than others. Thus, if a house owner wants his/her property to sell well, he/she should better build a basement, or have the house renovated.

Another interesting result is that the regression coefficients of floors and sqft_lot15_1, sqft_lot15_2, sqft_lot15_3, sqft_lot15_4 are all negative, meaning given all other conditions, houses with less floor levels, or less lot areas, are more expensive than others. This result implies that people have a preference for houses with less floor levels. As for the coefficients of sqft_lot, a possible explanation is that houses with big lot areas are usually located in remote districts. However, most people are still inclined to live in urban areas, where houses usually have small lots. Thus, despite the huge lots they include, housing prices in remote districts are still lower than those in urban areas, given all other conditions. This finding implies that total area, especially lot area, is somewhat trivial. But the exact district the house is located in is vital to its price.

The last interesting result is related to yrs_present. We can see that the coefficients of yrs_present and its quadratic term are all positive, but the interaction term of Yrs_present◊Renovation_1 is negative. This is to say, when the house is renovated, its regression coefficients related to yrs_present are expressed as below.

$$log\ (price) \approx (-2.553 \times 10^{-3}) * yrs\_present + (2.351 \times 10^{-5}) * yrs\_present^{\wedge 2} \\ + other\ terms$$

When the house is renovated, the relationship between housing price and yrs_present is quadric. This means given all other conditions, old houses and newly built houses are more expensive than houses of “middle age”. Houses newly built have modern facilities and better construction quality, which is quite normal to sell at good prices. But why are houses having 100 years of history or longer are more expensive than houses built 50 years ago? Actually, ordinary houses will not exist for as long as 100 years. They will be demolished and rebuilt in the previous trades. Only luxurious houses like manors, mansions, or houses with special history can last that long, which usually carry some antiques and fine art works. After proper renovation, these luxurious old houses are usually worth more than those with less history, given all other conditions.

In the above discussions, we analyzed some important factors that affect total housing price in King County. It is equally important to study their relationship with average housing price (per square feet). Define average housing price as below.

$$avg\_price = \frac{price}{total\_area\_except\_lot} = \frac{price}{sqft\_living15 + sqft\_above + sqft\_basement}$$

When calculating the total area of the house, we exclude sqft_lot15 for two reasons. First, sqft_lot15 is almost uncorrelated with housing price. Second, the definition of average housing price is given as the total price divided by total housing area, which only includes areas within the house. Also, sqft_lot15 has a huge deviation and many outliers. Including it will lead to many confusions and inaccuracy in prediction.

In case we are discussing average housing price, all factors concerning areas should be excluded, namely, sqft_above, sqft_basement. There is only one exception: sqft_lot15. Since it is not a part of the total housing area, we still keep it for prediction accuracy. Other factors, including quadratic and interaction terms, remain unchanged. We still take log-transformation of avg_price to stabilize variance of avg_price, and use stepwise regression with bidirectional elimination to reduce overfitting and collinearity problems. New model (model C) and fitted coefficients are shown below.

$$\begin{aligned} log\ (avg\_price) \sim\ & lat + yrs\_present + yrs\_present^{\wedge 2} + bedrooms + bathrooms + floors \\ & + grade + view + sqft\_lot15 + renovation + yrs\_present \times renovation \\ & + waterfront \end{aligned}$$

Variable	Estimate		Variable	Estimate	
(Intercept)	4.158	***	sqft_lot151	-9.961*10^{-2}	***
lat	1.763*10^{-3}	***	sqft_lot152	-1.982*10^{-1}	***
yrs_present	2.492*10^{-3}	***	sqft_lot153	-1.868*10^{-1}	***
yrs_present$^{\wedge 2}$	1.175*10^{-5}	***	sqft_lot154	-1.889*10^{-1}	***
bedrooms	-4.140*10^{-2}	***	waterfront1	4.323*10^{-1}	***
bathrooms	8.277*10^{-3}	*	renovation1	3.112*10^{-1}	***
grade	1.113*10^{-1}	***	yrs_present*renovation1	-3.544*10^{-3}	***
view	4.884*10^{-2}	***			

Table 2: Regression coefficients on log(avg_price)

In this regression on avg_price, residuals look pretty similar to a normal distribution. Stepwise method excluded one variable: floors, meaning that the coefficient of floors does not pass significance test. All remaining regression coefficients are significant. The adjusted R^2 is 0.5328, meaning that this model can explain 53.28% housing price variation, which is acceptable since we excluded most variables related to area.

Among all coefficients, we can see that the coefficients of lat, bathrooms, grade, view, renovation_1, waterfront_1 are all positive, which are exactly the same as the previous case. These results show that the increase in these factors will not only raise total housing price, but will also increase average housing price (per square foot). Besides, we can see that the magnitudes of coefficients are almost the same as the previous case. Grade, renovation, and waterfront still have the biggest positive influence on average housing price (per square foot).

An interesting result in this regression is that the coefficient of bedrooms reverses to negative, meaning that given all other conditions, houses with more bedrooms have lower price per square foot. This might sound counter-intuitive, but reasonable. Because the number of bedrooms is an indicator of housing area. Houses with lots of bedrooms usually have huge indoor areas. Based on economic knowledge, we know that the relationship between total price and housing area is not linear, and bigger houses tend to have lower price per square foot.

The regression coefficients of sqft_lot15_1, sqft_lot15_2, sqft_lot15_3, sqft_lot15_4 are still negative. As we have discussed before, sqft_lot is an indicator of house location. Normally speaking, houses with huge lots are usually located in rural areas remote from the city. Thus, this result implies that given all other conditions, houses in remote districts have lower average price per square foot compared to those in urban areas.

For the coefficient of yrs_present, the result remains the same as before. When the house is renovated, its regression coefficients related to yrs_present are expressed as below.

$$log\left(avg_{price}\right) \approx (-1.052 \times 10^{-3}) * yrs\_present + (1.175 \times 10^{-5}) * \ yrs\_present^{\wedge 2} \\ + other\ terms$$

When the house is renovated, the relationship between housing price and yrs_present is quadric. This means given all other conditions, the average price per square foot for old houses and newly built houses are higher than houses of “middle age”.

To conclude, all findings in the regression on total housing price, except that of bedrooms, remain unchanged in the regression on average housing price per square foot.

Part IV: Regression by Region

Because there exists obvious fluctuation of regression coefficients across different zip codes, which mainly include Seattle and its vicinity areas. According to administrative division, we divide the region into three parts: central Seattle (zip code from 98142 to 98160 and has 6413 samples), vicinity of Seattle (zip code from 98161 to 98170 and has 2564 samples) and Bellevue (zip code from 98004 to 98008 and has 1407 samples).

	Central		Vicinity			
Variable	Seattle		of Seattle		Bellevue	
(Intercept)	10.888	***	10.946	***	11.872	***
sqft_above	0.000	***	0.000	***	0.000	***
lat	0.001	***	0.001	***	0.004	***
bedrooms	-0.016	***	-0.031	***	-0.015	
bathrooms	0.055	***	0.082	***	0.054	***
floors	-0.013	**	-0.018	.	-0.025	*
grade	0.222	***	0.213	***	0.148	***
yrs_present	0.006	***	0.000		-0.015	***
I(yrs_present^2)	0.000	***	0.000	***	0.000	***
renovation1	0.264	***	0.235	*	0.781	***
sqft_basement1	0.151	***	0.223	***	0.141	***
sqft_lot151	-0.035	***	0.029		0.061	
sqft_lot152	-0.154	***	-0.103	***	0.079	.
sqft_lot153	-0.172	***	-0.128	***	0.211	***
sqft_lot154	-0.035		-0.045		0.273	***
yrs_present *renovation1	-0.003	***	-0.003	*	-0.011	***

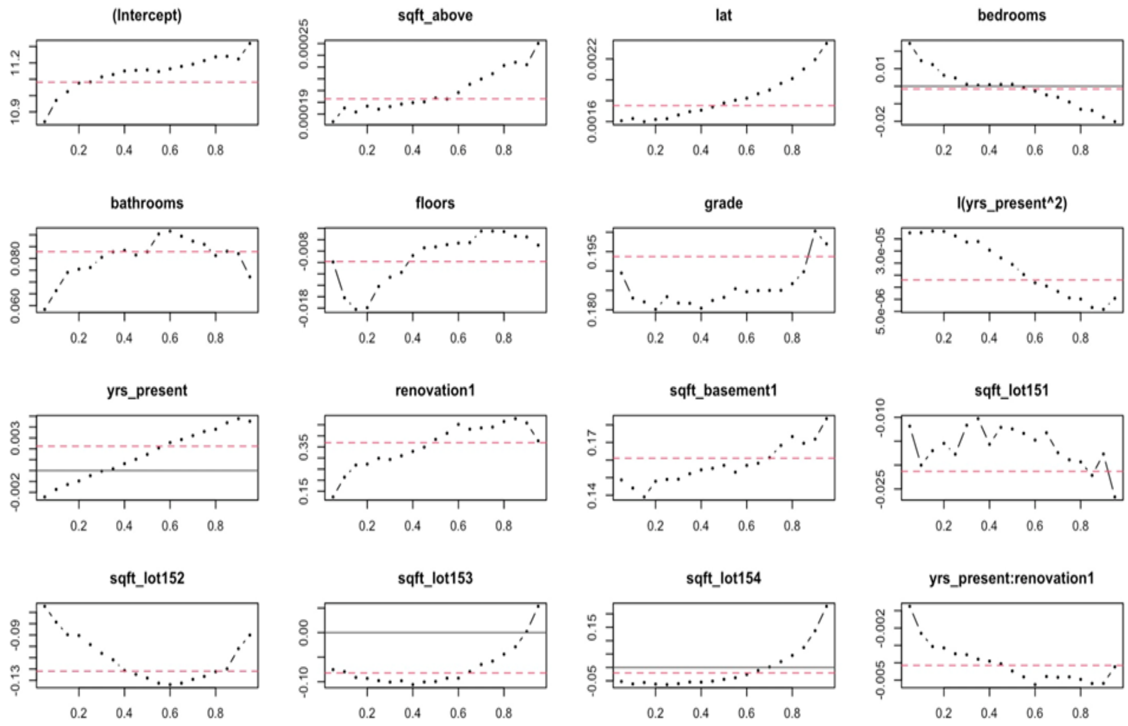
Table 3: Regression coefficients across different regions

As shown in table 3, the coefficient of yrs_present is positive in central Seattle while negative in Bellevue, which is different from original regression using all samples. Bellevue is a relatively new area compared to Seattle, so yrs_present is 41.556 on average, which is much less than that of 60.957 and 55.337 in central Seattle and the vicinity area of Seattle respectively. Therefore, the compensation of renovation for negative contribution to housing price caused by ages is not significant in Bellevue, which explains the negative coefficient of yrs_present in this area.

The coefficient of floors is more negative in Bellevue than that in Seattle area, because Bellevue is an upscale community with its location relatively independence of Seattle and very little crime, so many of middle and upper class in Seattle area moved to Bellevue in recent years, and it is reasonable that people lived in rural areas or wealthier people may have more preference for houses with less floors.

Part V: Quantile Regression

Quantile regression displays the variance of coefficients at different price quantiles. Select 20 quantiles with 0.05 interval between 0 and 1 and fit and smooth the model.



Plot 4: Coefficients in quantile regression

As shown in the Plot 4, for bedrooms, when the price is above median, bedrooms increasing comes with lower price, the higher the price is, the lower the coefficient is; when the price is below the median, bedrooms increasing comes with higher price, the higher the price is, the lower the coefficient is. This result is consistent with previous statements, increasing the number of bedrooms will gradually become a negative contribution to house price with the increase of the price quantile, that is, higher price houses usually have enough different types of rooms and will focus more on amenities and functionality rather than increasing the number of bedrooms.

For yrs_present, when the price above 30%: when the price is above median, years increasing comes with higher price, the higher the price is, the higher the coefficient is; when the price is below 30%, years increasing comes with lower price, the higher the price is, the less negative the coefficient is. For those higher priced houses, older houses imply better building materials or more renovations, which is a positive factor for housing price, while the negative contribution to housing price is not compensated by renovations for those lower price houses.

The coefficients for sqft_lot151 and sqft_lot152 (levels for smaller lot) are negative at each quantile, while coefficients for sqft_lot153 and sqft_lot1524(levels for larger lot) becomes positive at about 90% quantile and 70% quantile respectively. It implies that large lots may reduce the utilization of the land, which is the main concern of the lower price houses, henceforth reduce the house price, however, large lots may have positive contribution to those high-end houses.

Part VI: Conclusion

In this article, we primarily focus on discussing the factors that affect house price in King County. After processing the data into suitable structure, using the stepwise method, we fitted a linear regression model, with quadric and intersection incorporated, on logarithm of "price". The insertion of logarithm helps with avoiding problems like heteroskedasticity and residual non-normality. Concluding from our model, we have variables coefficients shown in table 1.

Then, we calculated the house price per square feet, which we note as variable "average price". We then similarly adopted the stepwise linear regression model with bidirectional elimination on logarithm of average price, with the variable to do with area, "sqft_above" and "sqft_basement", excluded in advance. By comparison with the model fitted for price, variable "floors" no longer kept significance for average price and the coefficient for "bedrooms" turned to negative. With 53.28% of the housing price variance can be explained, this model satisfies our expectation.

To move on, we further divided King County into different regions based on zip-code and performed regression analysis separately. With most variables' coefficients consistent with the model fitted for the whole King County, the magnitude of coefficient for "Yrs_present" and "floors" do vary among different regions. Our final step to polish our model is quantile regression of size 20. We found that most variables' coefficients held still among different price levels. But the tendency of effects of "yrs_present" and "bedroom" vary as house price falls in various levels. These further findings are illuminating and complete our analysis.

Acknowledgements

We really appreciate the instruction from professor Carlos Cinelli and our TA Kenny Zhang. Besides, this work could have not been completed without the data shared on Kaggle.

References

[1] Algieri, Bernardina. "House price determinants: Fundamentals and underlying factors." *Comparative Economic Studies* 55.2 (2013): 315-341.

[2] Capozza, Dennis R., et al. "Determinants of real house price dynamics." (2002). *NBER*.

[3] Rawat, Neeraj Singh. “Data Analysis with Python: House Sales in King County, USA”

[4] Untitled. “House Sales in King County, USA”. *Kaggle*.

[5] Untitled. “Residential Glossary of Terms”. *King County*.

Appendix I: Variables discriptions

id:	A notation for a house
date:	Date house was sold
price:	Price of house at sale
bedrooms:	Number of bedrooms
bathrooms:	Number of bathrooms
sqft_living:	Square footage of the home
sqft_lot:	Square footage of the lot
floors:	Total floors (levels) in house
waterfront:	House which has a view to a waterfront
view:	Overall grade of the view
condition:	How good the condition is overall
grade:	Overall grade given to the housing unit, based on King County grading system[6]
sqft_above:	Square footage of house apart from basement
sqft_basement:	Square footage of the basement
yr_built:	Built Year
yr_renovated:	Year when house was renovated
zip code:	Zip code
lat:	Latitude coordinate
long:	Longitude coordinate
sqft_living15:	Living room area in 2015(implies-- some renovations)
sqft_lot15:	LotSize area in 2015(implies-- some renovations)
yr_recorded:	The year when the sale is recorded
yrs_present:	The number of years the house has existed until sale
total_area:	Total area of the house including above, living, basement, and lot
total_area_except_lot:	Total area of the house including above, living, and basement
renovation:	Whether the house has been renovated

[6] See more for definitions of grade, etc. Untitled. "Residential Glossary of Terms". *King County*. https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r

Appendix II: Variables types and transformations

Type	Variable	Transformation
Continuous	Sqft_above	
	Lat	Lat=(Lat-E(Lat))*100/SD(Lat)
	Yrs_present	
Categorical, but viewed as continuous	Bedrooms	
	Bathrooms	

	Floors	Floors=1, if floors=1.0
		Floors=2, if floors=1.5
		Floors=3, if floors=2.0
		Floors=4, if floors=2.5
		Floors=5, if floors=3.0
		Floors=6, if floors=3.5

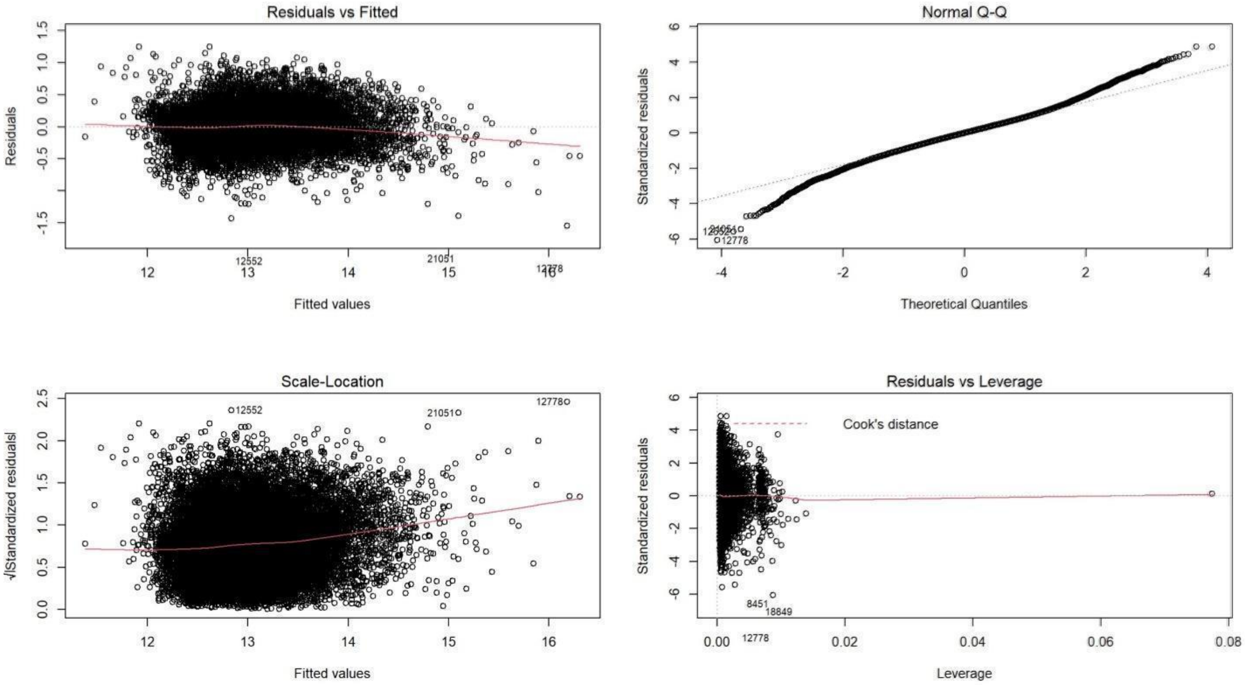
	Grade	
	View	

Categorical	Sqft_basement	Sqft_basement=0, if no basement
		Sqft_basement=1, if exist basement

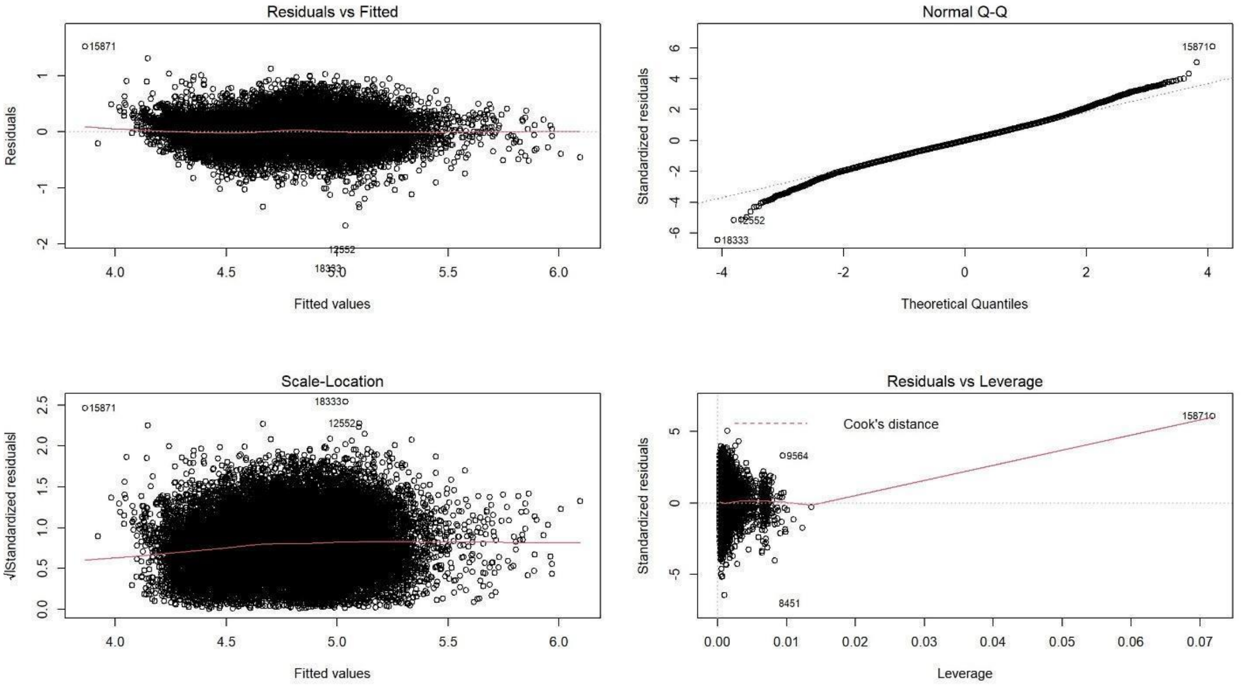
	Sqft_lot15	Sqft_lot15=0, if between (651,4800)
		Sqft_lot15=1, if between (4800,6750)
		Sqft_lot15=2, if between (6750,8382)
		Sqft_lot15=3, if between (8382,11094)
		Sqft_lot15=4, if between (11094,871200)

	Renovation	
	Waterfront	

Appendix III: Diagnosis plots of regression



Appendix III plot 1: diagnosis plots of regression on log(price)



Appendix III plot 2: diagnosis plots of regression on log(avg_price)