Analysis of real estate unit prices

Introduction

This project aims to analyze the average pattern of association between the unit price of houses and the size (living area) of the house, and reveal additional key attributes of the house that influences this relationship. With this project I aim to gain deeper understanding on the real estate market's pricing mechanisms, and see what attributes of a house elevates its market value.

My research question is: What is the average pattern of association between the area and the unit price of a house, and what are the additional important factors that influence this relationship?

My initial hypothesis is that unit price decreases by size, and given two houses with the same size, the key differentiating factors are the lot size, the quality of the building, and the neighborhood.

My left-hand side (Y) variable is the unit price of houses, in USD per square feet. My main right hand side variable is the area of the house, in square feet. I considered the log transformation of these variables and I used additional confoundres (attributes of houses) to reveal the relationship.

Data

To analyze this question, I have obtained a dataset on houses sold in King County, USA, from 2014.05 to 2015.05.

Datasource: The Center for Spatial Data Science, University of Chicago https://geodacenter.github.io/data-and-lab//KingCounty-HouseSales2015/

The dataset contains 21,613 observations, and 21 variables. The description of all the variables can be found in the Appendix.

The data was obtained already in a cleaned format. I have excluded observations with no bedrooms, bathrooms, and with lots that are 100 times larger than the house. These observations are likely to not describe houses but other buildings or just the lot.

The model should reach high external validity, as I have excluded variables which are specific to King County: ZIP code, latitude, longitude, average values of 15 nearest houses.

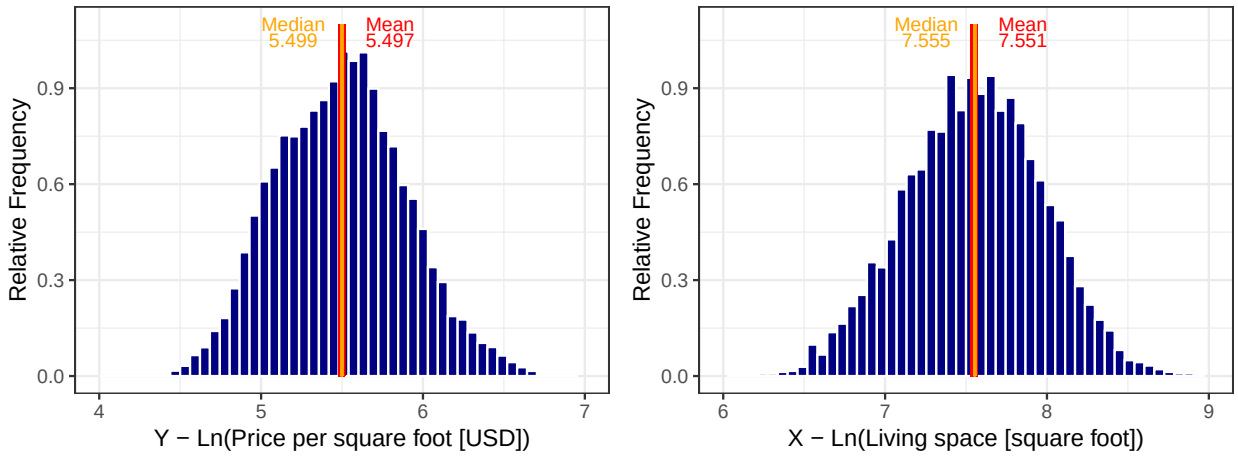
I modified the remaining variables to prepare the data set for the analysis. The key variables and modifications I made can be found in the table below, and the distributions of the transformed variables can be found in the Appendix.:

Variable	Description	Transformation
ln_pps	The natural logarithm of the ratio of price and living area (Y variable)	Natural logarithm, to shift skewed distribution to normal, and to be able to talk about differences
ln_living	The natural logarithm of size of the house's living area in square feet	Natural logarithm, to to shift skewed distribution to normal, and to be able to talk about differences
grade2	Classification by construction quality, referring to types of materials and workmanship from 1 to 7	Condensing original 11 bins into 7 bins to avoid bind with few (<100) observations
ln_lot	The natural logarithm of the lot area	Natural logarithm, to shift skewed distribution to normal, and to be able to talk about differences
age	Time passed in years between the construction year and present (2015)	Transforming year built to age
bathrooms2	Number of bathrooms from 1 to 5+	Condensing actual number of bathrooms to bins 1-5+, due to too few observations (>100) in some binss
bedrooms2	Number of bedrooms from 1 to 6+	Condensing actual number of bedrooms to bins 1-6+, due to too few observations (>100) in some binss
condition2	Condition of the house, ranked from 1 to 5	Condensing to bins 1-4, due to too few observations (>100) in some bins
view2	An index from 0 to 2 of how good the view of the property was	Condensing to bins 0-2, due to too few observations (>100) in some bins
waterfront	'1' if the property has a waterfront, '0' if not	No transformation
renovation_bin	Renovation status of the building, with renovated since 2000, since 1980, or before 1980 or never	Creating renovation bin from year of renovation

In the raw dataset, both price and living area had a log distribution, as can be seen in the Appendix. Thus their ratio, the price per square feet, was skewed to the right (Y variable), . This was also true for my X variable, living area.

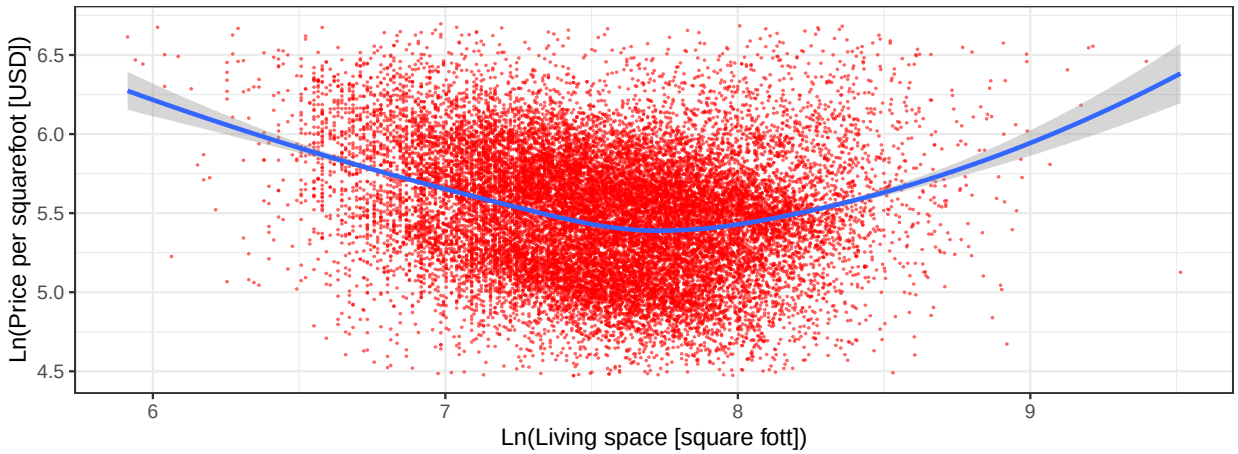
The log transformation of Y and X ensured that these variables are normally distributed. The interpretation of the results are also more favorable this way, by being able to talk about price differences in realtive terms.

The mean and the median are very close to each other for both variables as can be seen below:



The scatter plot below of X and Y variables reveals that the average pattern of association suggests larger houses have a lower unit price, but only up to a certain point. Near log(x)=7.7 the loess slope changes from negative to positive. Meaning that for large houses, my initial hypotheses could be false.

In the model, I will include ln_living as a linear spline with a knot at 7.7. I will test whether the two segments really have a different slope.



To better explain the relationship between Y and X, I considered to include confounders.

I chose how to model each confounder as described in the table below. Figures of the relationship between the confounderes and Y can be found in the Appendix to support my decisions.

Variable	Modeling.method
grade2	Include as a dummy variable to capture the non-linear increase of means per grades
ln_lot	Include as linear spline with a knot at 9.0
age	Loess suggests a third-order relationship, but for the sake of better interpretation I wil use linear spline with a knot at 25
bathrooms2	Include as a dummy variable to capture the increasing differences of means per number of bathrooms
bedrooms2	Include as a dummy variable to capture the de/increasing differences of means per number of bedrooms
condition2	Include as a dummy variable to capture the non-uniformly increasing differences of means per condition level
view2	Include as a dummy variable to capture the increasing differences of means per view level
waterfront	Include as a binry dummy variable
renovation	Include as a dummy variable with 3 levels

Model

	reg1	reg2	reg3	reg4	reg5	reg6	reg_all
Constant	8.720**	10.181**	10.833**	10.162**	10.190**	10.301**	9.931**
	(0.075)	(0.086)	(0.088)	(0.084)	(0.083)	(0.088)	(0.100)
ln_living<7.7	−0.438**	−0.709**	−0.618**	−0.574**	−0.580**	−0.608**	−0.556**
	(0.010)	(0.011)	(0.012)	(0.011)	(0.011)	(0.012)	(0.014)
ln_living>7.7	0.339**	−0.280**	−0.209**	−0.255**	−0.304**	−0.382**	−0.359**
	(0.017)	(0.021)	(0.021)	(0.020)	(0.019)	(0.021)	(0.022)
f(grade=2)		0.165**	0.144**	0.177**	0.182**	0.185**	0.187**
		(0.028)	(0.028)	(0.029)	(0.029)	(0.029)	(0.028)
f(grade=3)		0.352**	0.297**	0.433**	0.436**	0.435**	0.439**
		(0.028)	(0.028)	(0.029)	(0.028)	(0.028)	(0.028)
f(grade=4)		0.561**	0.469**	0.671**	0.665**	0.665**	0.662**
		(0.028)	(0.029)	(0.030)	(0.029)	(0.029)	(0.029)
f(grade=5)		0.766**	0.679**	0.912**	0.901**	0.907**	0.899**
		(0.029)	(0.030)	(0.030)	(0.030)	(0.030)	(0.029)
f(grade=6)		0.943**	0.866**	1.113**	1.095**	1.095**	1.084**
		(0.031)	(0.031)	(0.032)	(0.031)	(0.031)	(0.031)
f(grade=7)		1.172**	1.088**	1.352**	1.307**	1.291**	1.278**
		(0.035)	(0.035)	(0.035)	(0.034)	(0.034)	(0.034)
f(condition=2)		0.091**	0.055	0.143**	0.142**	0.143**	0.140**
		(0.034)	(0.033)	(0.032)	(0.032)	(0.031)	(0.031)
f(condition=3)		0.179**	0.172**	0.181**	0.180**	0.179**	0.181**
		(0.034)	(0.033)	(0.032)	(0.032)	(0.032)	(0.032)
f(condition=4)		0.322**	0.297**	0.236**	0.233**	0.229**	0.233**
		(0.035)	(0.034)	(0.033)	(0.032)	(0.032)	(0.032)
ln_lot<9			−0.142**	−0.146**	−0.143**	−0.135**	−0.127**
			(0.005)	(0.005)	(0.005)	(0.006)	(0.006)
ln_lot>9			−0.010*	0.011**	0.008*	0.010*	0.006
			(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
age<25				0.000	−0.000	0.000	0.000
				(0.000)	(0.000)	(0.000)	(0.000)
age>25				0.006**	0.006**	0.006**	0.006**
				(0.000)	(0.000)	(0.000)	(0.000)
f(view=1)					0.106**	0.103**	0.098**
					(0.009)	(0.009)	(0.009)
f(view=2)					0.264**	0.261**	0.255**
					(0.024)	(0.024)	(0.023)
waterfront					0.356**	0.354**	0.341**
					(0.034)	(0.034)	(0.034)
f(bathrooms=2)						0.032**	0.037**
						(0.008)	(0.008)
f(bathrooms=3)						0.070**	0.079**
						(0.011)	(0.011)
f(bathrooms=4)						0.122**	0.134**
						(0.014)	(0.014)
f(bathrooms=5+)						0.235**	0.264**
						(0.049)	(0.049)
f(bedrooms=2)							−0.030
							(0.032)
f(bedrooms=3)							−0.092**
							(0.032)
f(bedrooms=4)							−0.112**
							(0.032)
f(bedrooms=5)							−0.122**
							(0.033)
f(bedrooms=6+)							−0.183**
							(0.038)
Renovated: 1980-2000							0.002
							(0.019)
Renovated since 2000							0.070**
							(0.017)
Num.Obs.	21 447	21 447	21 447	21 447	21 447	21 447	21 447
R2	0.086	0.235	0.261	0.379	0.402	0.404	0.408
Std.Errors	HC1	HC1	HC1	HC1	HC1	HC1	HC1

* p < 0.05, ** p < 0.01

I included 5 models, adding increasingly more variables, and a final model that contains all variables. The result of my models are summarized in the table above.

Holistic description of model results:

- In all models the slope of the living area (X variable) in both the two linear spline segments (>7.7 and <7.7) are both statistically significant and different from each other, as the CI intervals do not overlap -By adding more confounders, the slope of the <7.7 segment decreases, and the slope of the >7.7 segment increases
- The R square increases from 0.086 by adding more and more confounders, however it only reaches 0.408 by adding all variables. This means the model only explains c.41% of the variation, and there is still a significant amount of noise included
- Adding number of bedrooms and renovation status to Regression 6 increased the R square from 0.405 to 0.408 only, and decreases the slope of the living area (X variable) in the <7.7 segment
- Renovation stat

My preferred model is Regression 7, from which we can draw the following conclusions:

- The higher the living area, the lower the unit price is on average
- For houses with a log living area below 7.7, houses with a living area larger by 1% have 0.61% lower unit price on average, (given that allo other variables are the same)
- For houses with a log living area above 7.7, this relationship is weaker: houses with a living area larger by 1% have 0.38% lower unit price on average, given that allo other variables are the same.
- This can be explained with large houses being more high-end, where the quality and other external attributes have a higher importance
- Two variables are omitted: number of bedrooms and renovation status

Interpretation of grade: - The higher the grade, the higher the unit price, e.g. a grade 7 house being 1.29 times more expensive compared to a grade 1 house on average, (given that all other variables are the same - All grade levels are statistically significant

Interpretation of condition: - The better the condition, the higher the unit price, e.g. a condition 4 house having a 23% higher price compared to a condition 1 house on average, given that all other variables are the same - All condition levels are statistically significant

Interpretation of lot size: - For houses with log lot size smaller than 9.0, houses with a 1% larger lot size have 0.135% higer unit price on average, given that all other variables are the same - For houses with log lot size larger than 9.0, houses with a 1% larger lot size have 0.01% higer unit price on average, given that all other variables are the same - All condition levels are statistically significant

Interpretation of age: - Below age 25, the result is statistically not significant, meaning age has no effect below 25 years - Houses older than 25 years have a higher unit price, with 1 year older house being 0.6% more expensive on average, given that all other variables are the same

Interpretation of view: - Houses with a better view have higher unit price on average, .e.g. a view level 2 house has 26% higher unit price on average compared to a house with no view (view level 0), given that all other variables are the same

Interpretation of waterfront: - Houses with access to the water have a 35% higer unit price on average, than houses with no waterfront access, given that all other variables are the same

Interpretation of bathrooms: - Houses with more bathrooms have a hiher unit price, e.g. a house with 5+ bathrooms have a 24% higher unit price on average than houses with 1 bathrooms, given that all other variables are the same - The increase in unit price per one more bathroom is close to exponential

External validity

The model has high external validity in time.

To check this, I split the dataset into two parts to check external validity in time. I inclded 5000 observations into a dataset starting from the first observation's date (2014-05 month) and I included 5000 observations into another dataset going back from the last observation's date (2015-05).

From the results we can conclude that the model has high external validity in time: even though the coefficients are somewhat different, they are not significantly different.

```
## [1] "double"
```

	reg-2014-May-June	reg-2015-Feb-May
Constant	10.036**	10.392**
	(0.180)	(0.185)
ln_living<7.7	−0.604**	−0.581**
	(0.026)	(0.026)
ln_living>7.7	−0.378**	−0.422**
	(0.043)	(0.046)
f(grade=2)	0.235**	0.152*
	(0.075)	(0.060)
f(grade=3)	0.470**	0.408**
	(0.074)	(0.059)
f(grade=4)	0.703**	0.626**
	(0.075)	(0.061)
f(grade=5)	0.956**	0.869**
	(0.076)	(0.062)
f(grade=6)	1.115**	1.060**
	(0.078)	(0.065)
f(grade=7)	1.286**	1.266**
	(0.083)	(0.072)
f(condition=2)	0.244**	0.068
	(0.056)	(0.064)
f(condition=3)	0.282**	0.101
	(0.056)	(0.064)
f(condition=4)	0.358**	0.156*
	(0.057)	(0.066)
ln_lot<9	−0.127**	−0.151**
	(0.011)	(0.012)
ln_lot>9	−0.002	0.027**
	(0.009)	(0.008)
age<25	0.001	0.000
	(0.001)	(0.001)
age>25	0.006**	0.007**
	(0.000)	(0.000)
f(view=1)	0.071**	0.113**
	(0.019)	(0.019)
f(view=2)	0.183**	0.320**
	(0.041)	(0.053)
waterfront	0.420**	0.273**
	(0.060)	(0.085)
f(bathrooms=2)	0.059**	0.008
	(0.018)	(0.017)
f(bathrooms=3)	0.096**	0.045*
	(0.022)	(0.023)
f(bathrooms=4)	0.133**	0.141**
	(0.029)	(0.029)
f(bathrooms=5+)	0.201*	0.258
	(0.083)	(0.133)
Num.Obs.	5000	5000
R2	0.399	0.406
Std.Errors	HC1	HC1

* p < 0.05, ** p < 0.01

Conclusion

To anwser my research question: Houses with alrger living area are indeed cheaper. This is more true for houses with a log living area below 7.7.

A key confounder is grade, determining the quality of materials and workmanship. Meaning that in case of two houses with same living area, the grade will be the key differentiating factor in their unit prices.

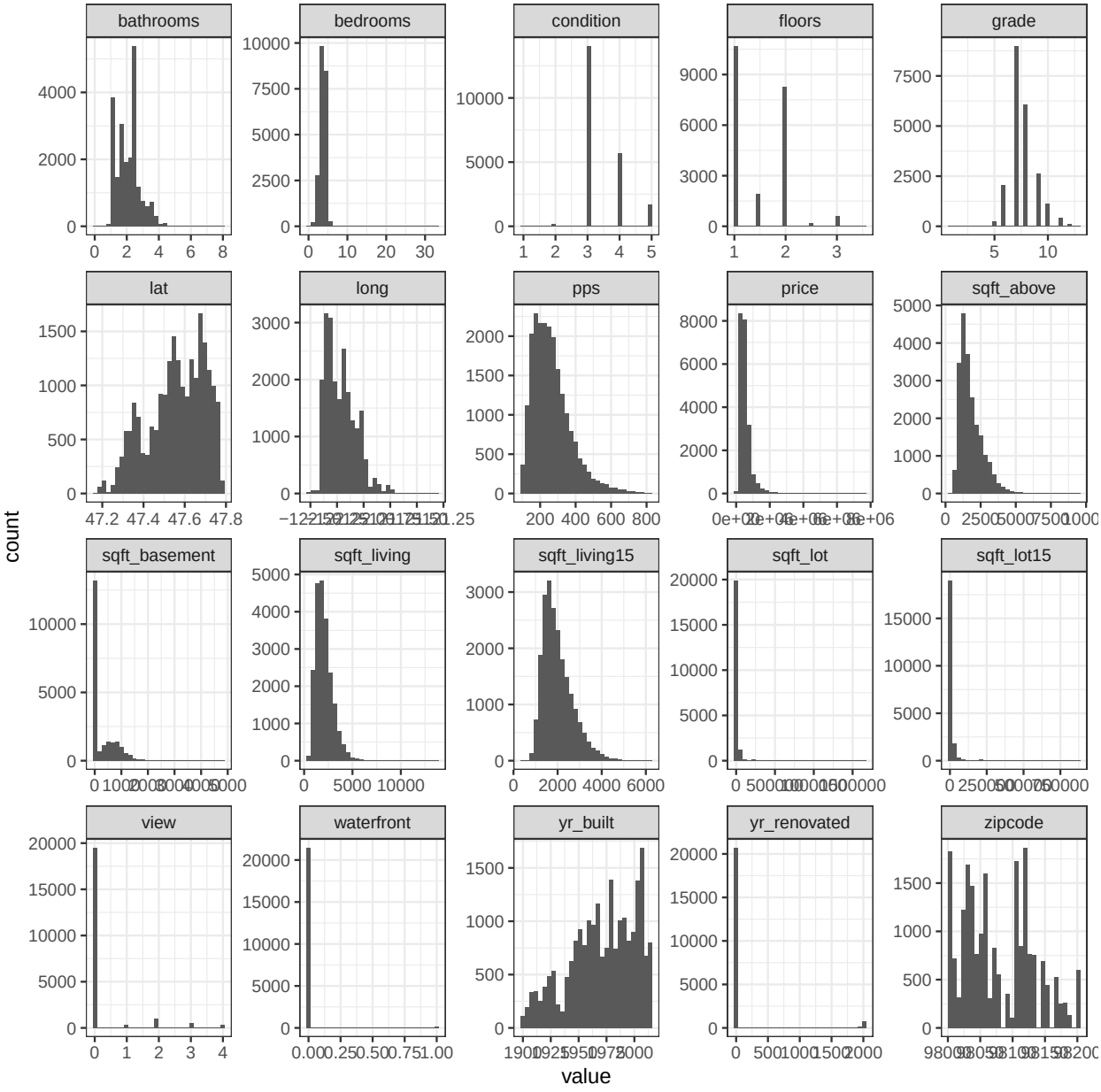
Additional important confounders are: waterfront, number of bathrooms, view, condition. Suprizingly, lot size is not as important as I initially considered it to be.

Appendix

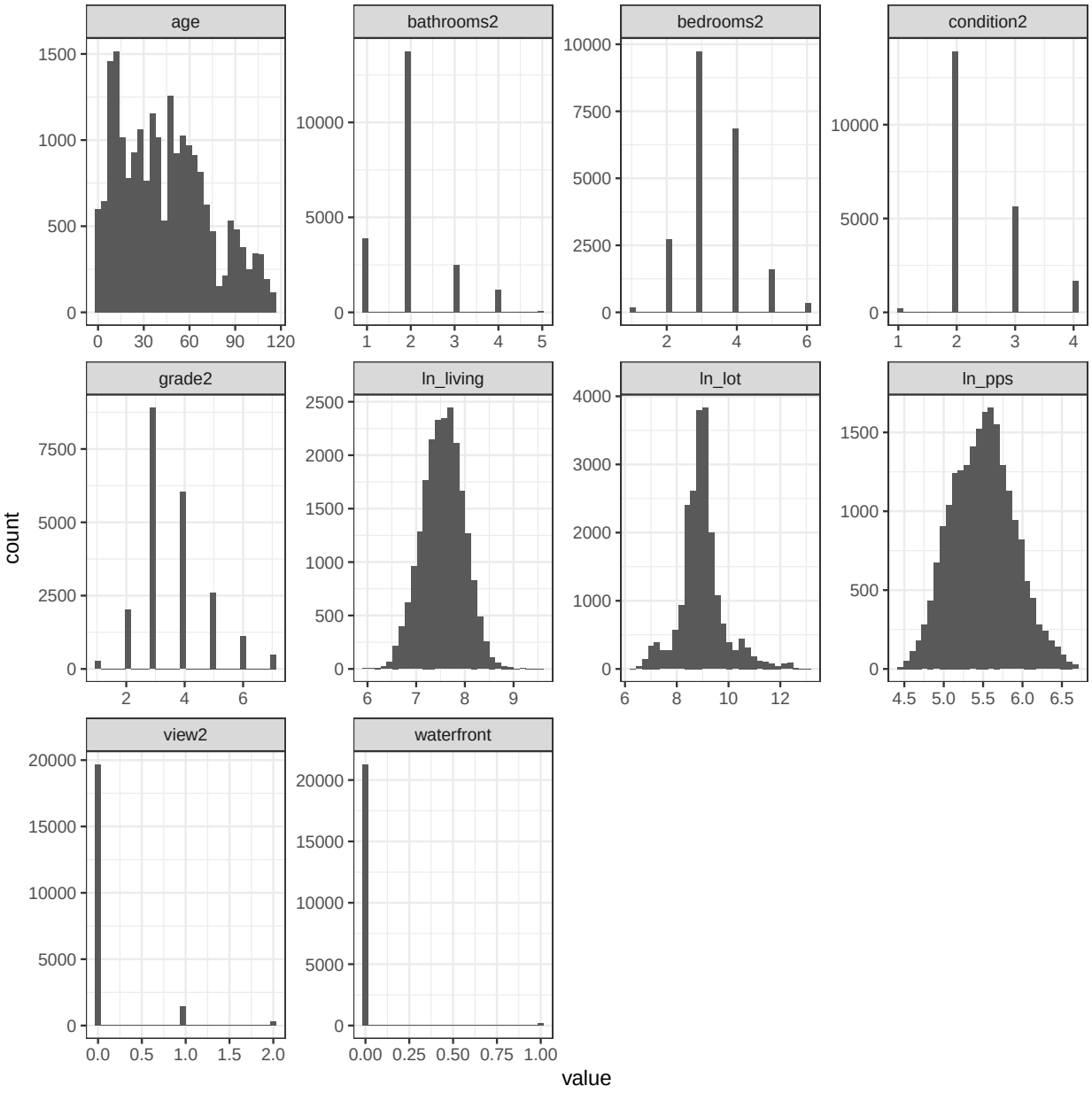
Description of the original dataset:

Variable	Description
id	Identification
date	Date sold
price	Sale price [USD]
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
sqft_liv	Size of living area in square feet
sqft_lot	Size of the lot in square feet
floors	Number of floors
waterfront	'1' if the property has a waterfront, '0' if not.
view	An index from 0 to 4 of how good the view of the property was
condition	Condition of the house, ranked from 1 to 5
grade	Classification by construction quality which refers to the types of materials used and the quality of workmanship
sqft_above	Square feet above ground
sqft_basmt	Square feet below ground
yr_built	Year built
yr_renov	Year renovated. '0' if never renovated
zipcode	5 digit zip code
lat	Latitude
long	Longitude
squft_liv15	Average size of interior housing living space for the closest 15 houses, in square feet
squft_lot15	Average size of land lots for the closest 15 houses, in square feet
Shape_leng	Polygon length in meters
Shape_Area	Polygon area in meters

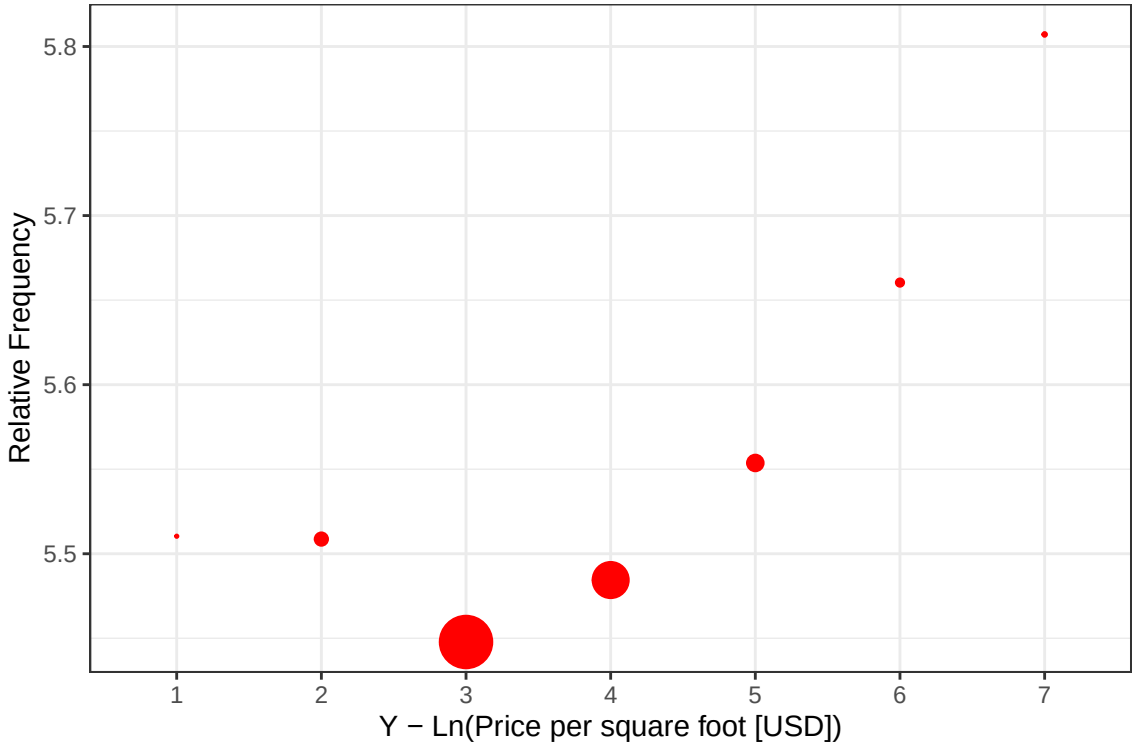
Distributions of the variables in the original dataset:



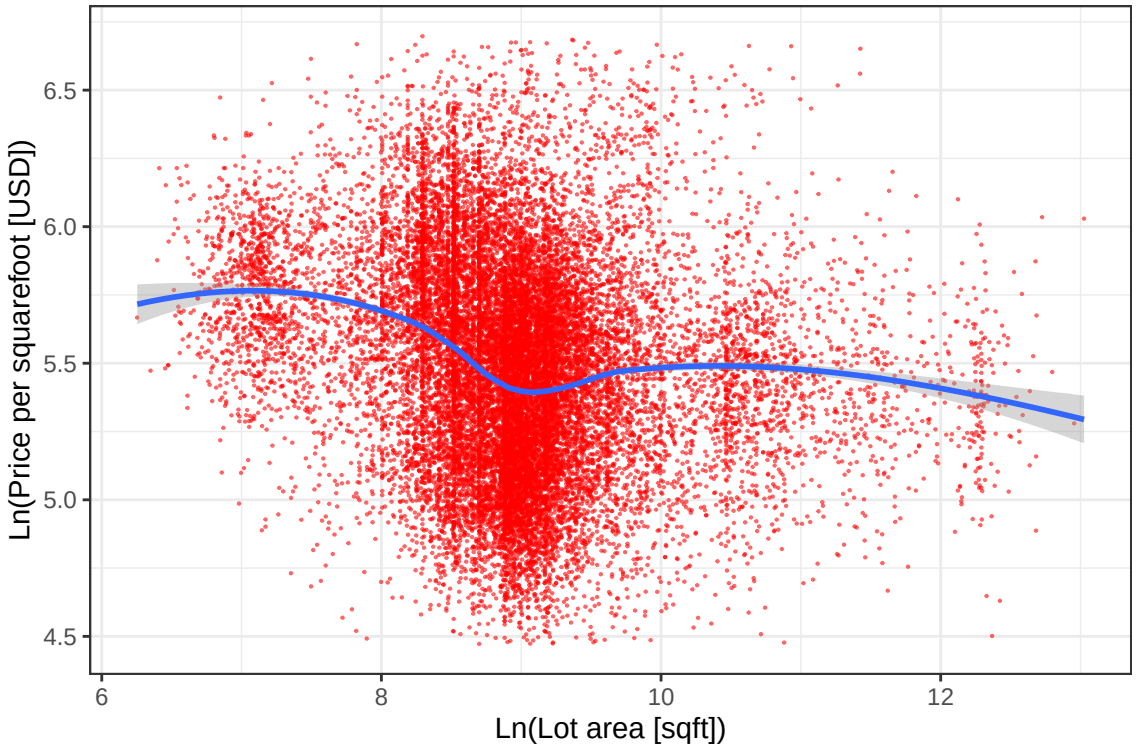
Distributions of the variables in the transformed dataset:



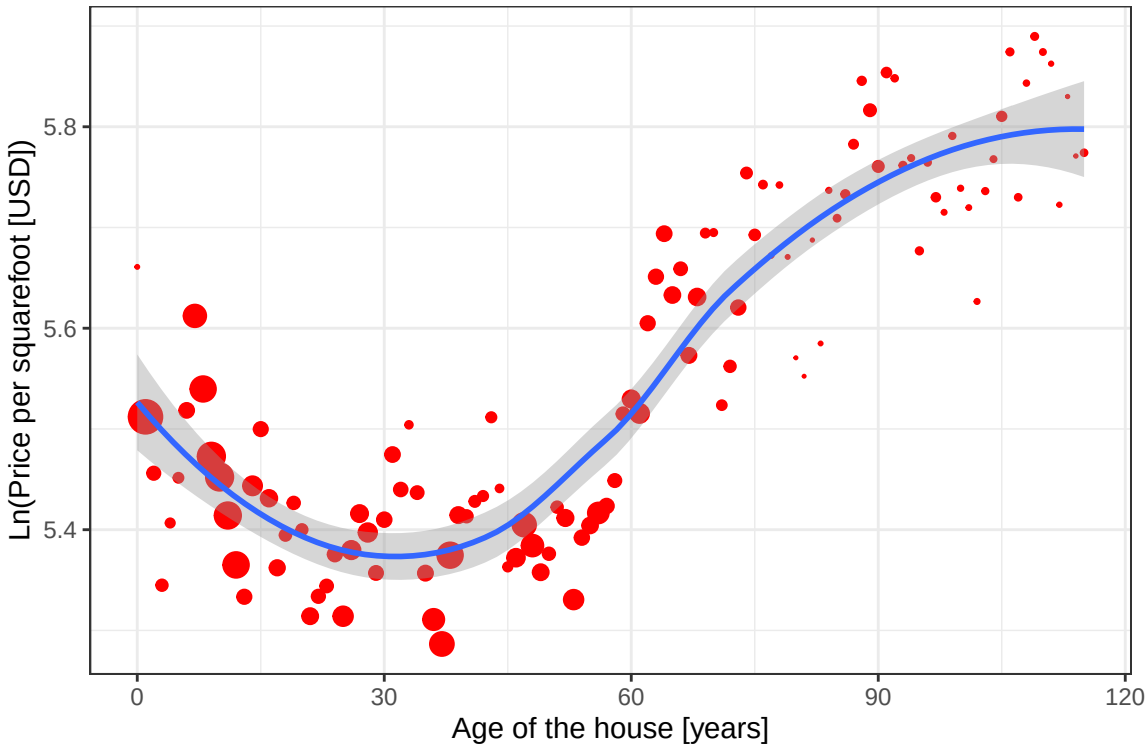
Grade2 and Unit price relation:



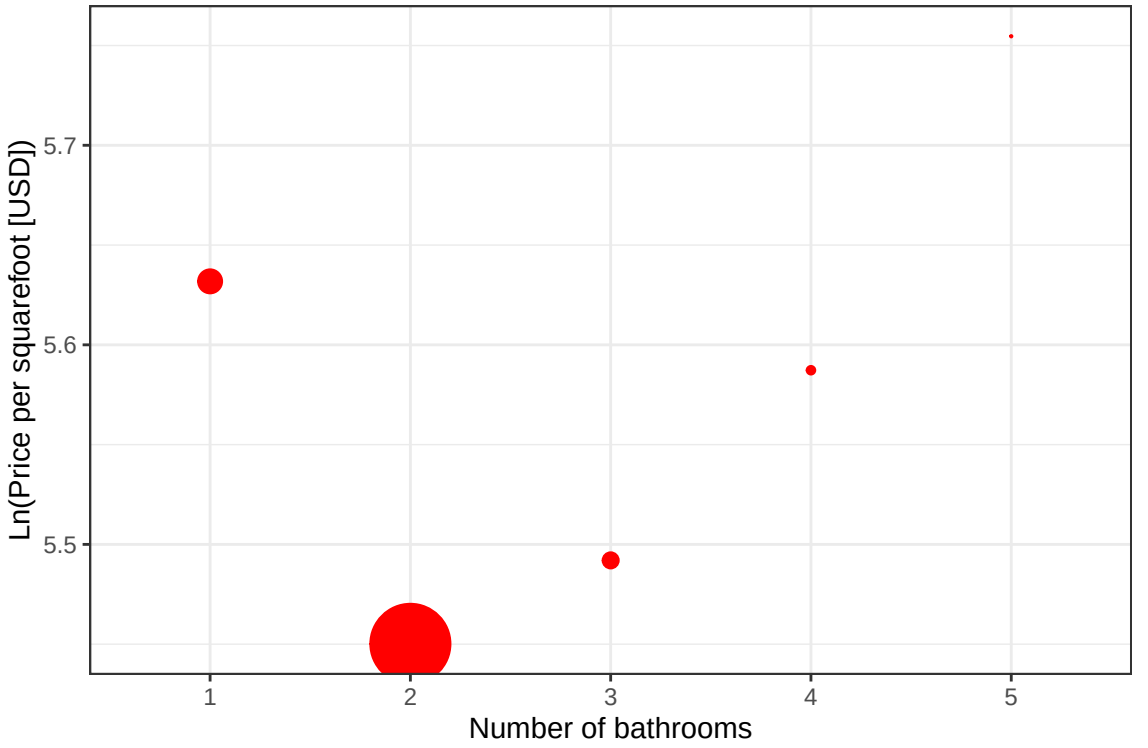
Lot size and Unit price relation:



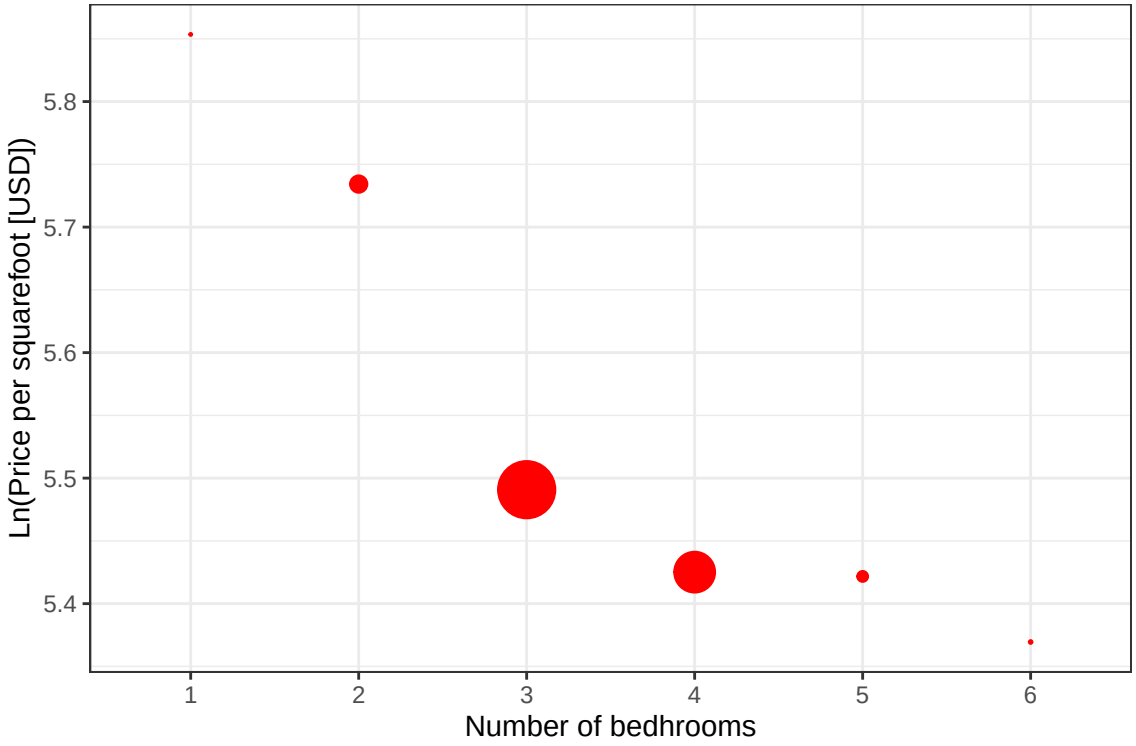
Age and Unit price relation:



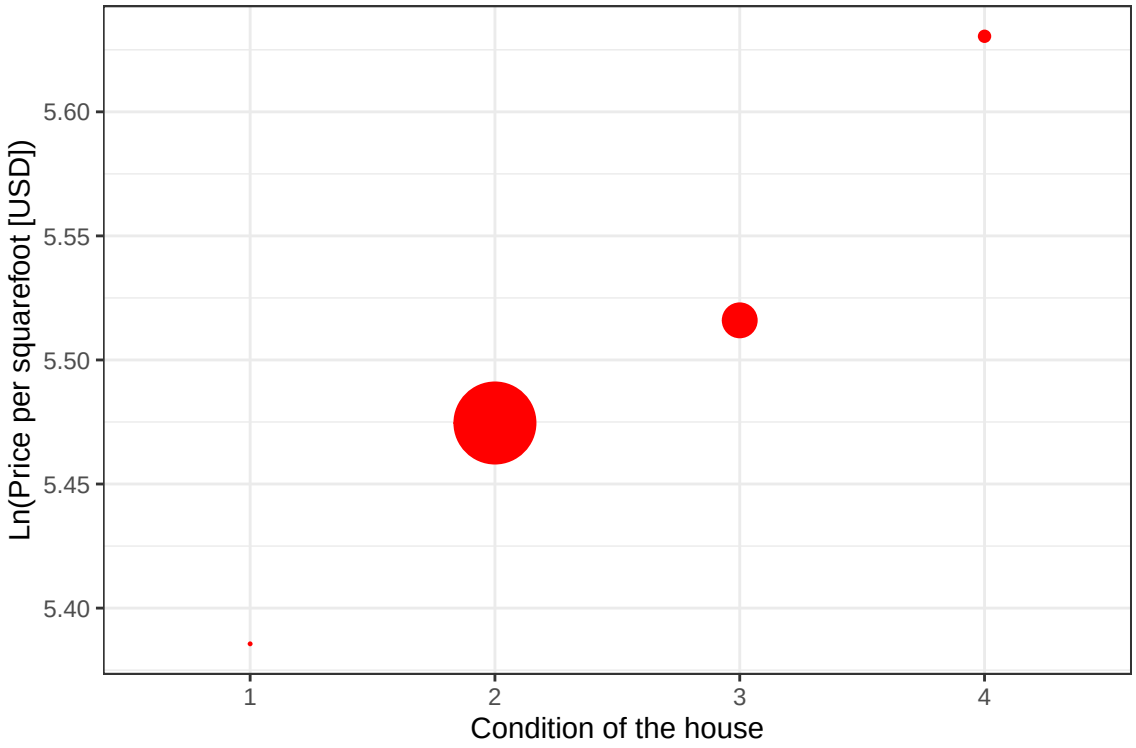
Bathroom number and Unit price relation:



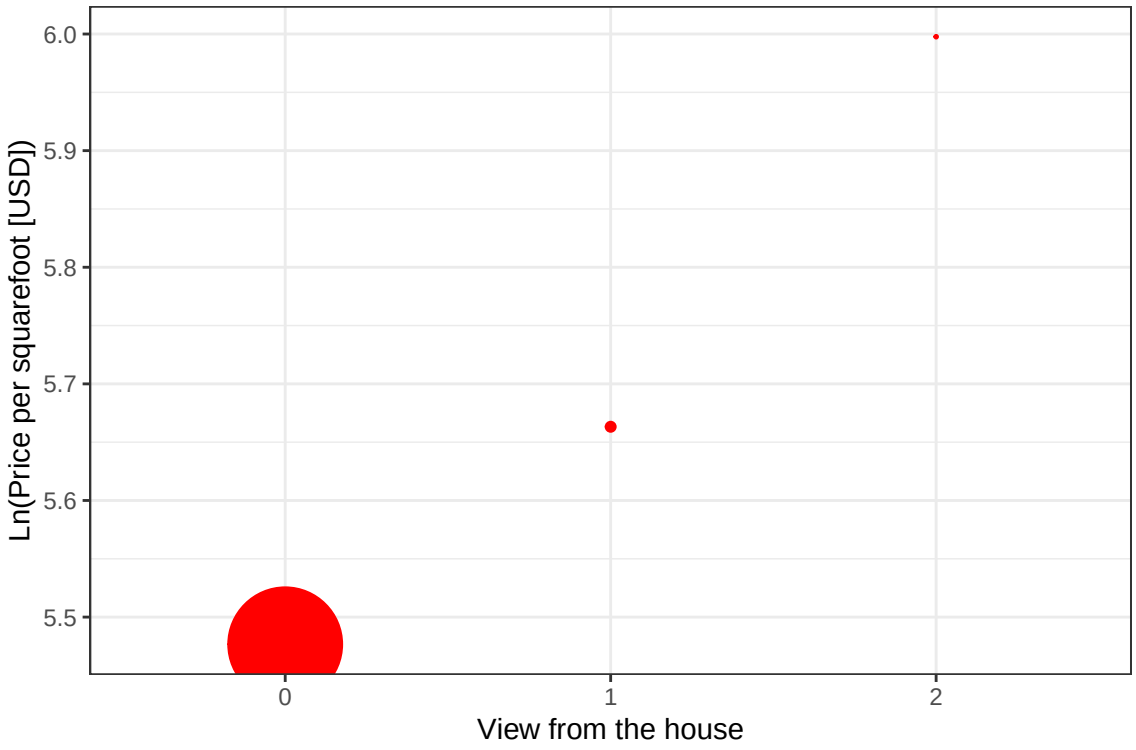
Bedroom number and Unit price relation:



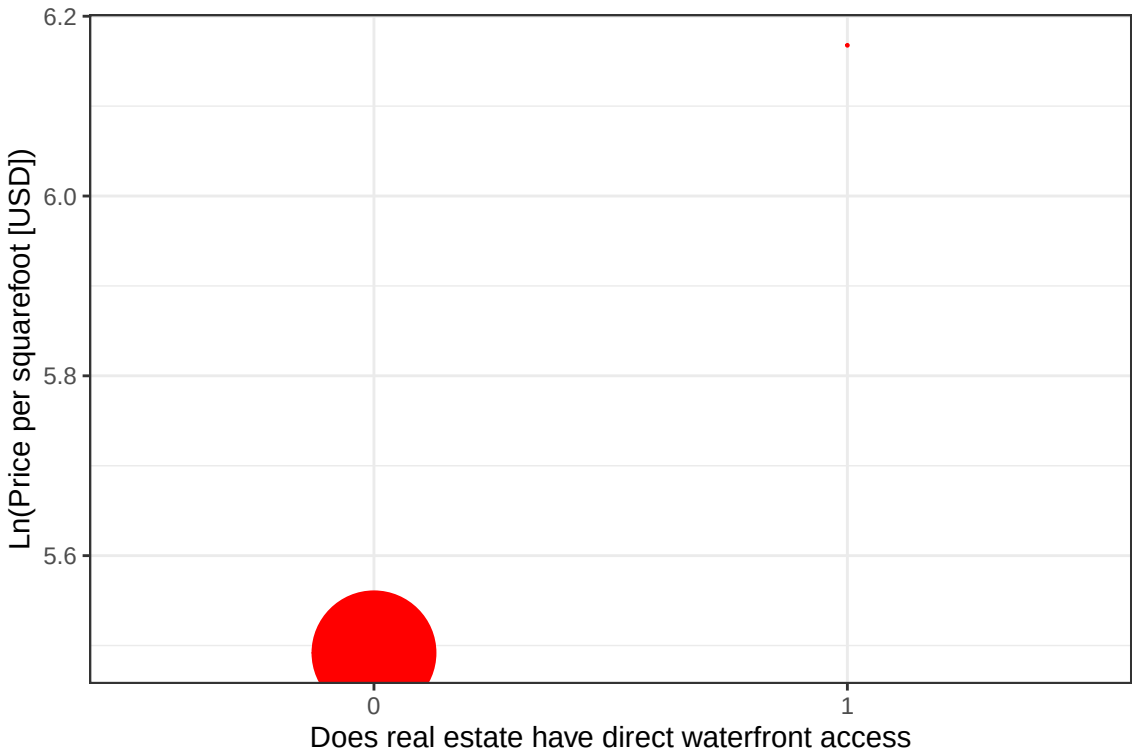
Condition2 and Unit price relation:



View2 and Unit Price relation:



Waterfront and Unit Price relation:



Renovation and Unit Price relation: